ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

ESCUELA DE FÍSICA Y MATEMÁTICA

CARACTERIZACIÓN FENOTÍPICA Y GENÉTICA DE CLONES PROMISORIOS DE CACAO (*THEOBROMA CACAO L.*) CON ANÁLISIS MULTIVARIANTE PROCRUSTES

TRABAJO DE TITULACIÓN

TIPO: PROYECTO DE INVESTIGACIÓN

Presentado previo a la obtención del título de:

INGENIERA EN ESTADÍSTICA INFORMÁTICA

AUTORA: GABRIELA JOSSELYN OBREGÓN ORTIZ

TUTOR: DR. RUBÉN ANTONIO PAZMIÑO MAJI

Riobamba-Ecuador

2018

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

ESCUELA DE FÍSICA Y MATEMÁTICA

El Tribunal del Trabajo de titulación certifica que: El trabajo de investigación: Caracterización fenotípica y genética de clones promisorios de cacao (*Theobroma cacao L.*) con análisis multivariante procrustes, de responsabilidad de la señorita Gabriela Josselyn Obregón Ortiz, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de titulación, quedando autorizada su presentación.

	FIRMA	**FECHA**
Dr. Rubén Antonio Pazmiño Maji		
DIRECTOR DE TESIS	_____	_____
Ing. Pablo Javier Flores Muñoz		
MIEMBRO DEL TRIBUNAL	_____	_____

Yo, Gabriela Josselyn Obregón Ortiz declaro que el presente trabajo de titulación es de mi autoría y que los resultados del mismo son auténticos y originales. Los textos constantes en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autora asumo la responsabilidad legal y académica de los contenidos de este trabajo de titulación.

Gabriela Josselyn Obregón Ortiz

DEDICATORIA

A los seres que por medio del amor me acompañan durante esta existencia:

Gabriel, Yara, Carolina, Jennifer, Miguel.

Gabriela

AGRADECIMIENTO

Quiero manifestar mi más grande agradecimiento a todos quienes aportaron de una u otra forma con la realización de este proyecto:

En primer lugar, a mi padre Gabriel Obregón, quien además de financiar por completo este trabajo, me demostró su amor de todas las formas posibles, por medio de su empuje, paciencia y motivación, brindándome todo el apoyo logístico, viajando, e incluso desvelándose conmigo durante el proceso de recolección de datos.

A Fernando Romero, primer director, quien ideó este trabajo y tuvo la visión de desarrollar este análisis en este importante cultivo, por proporcionarme el tema de tesis, orientarme, saber llevarme donde los contactos adecuados, viajar conmigo, conseguir el análisis molecular, y su muy buena voluntad y ganas de aportar.

A Rubén Pazmiño, por dirigir este proyecto brindando su valiosa guía en el análisis estadístico y en la estructura del trabajo escrito.

A Pablo Flores, por las correcciones y sugerencias para mejorar este trabajo.

Al Programa de Cacao de la Estación Experimental Litoral Sur del INIAP, en las personas de:

James Quiroz, director del Programa de Cacao, por su gran apertura al brindarme la oportunidad y las facilidades para desarrollar este proyecto en la Estación, además de su valioso asesoramiento, sugerencias y colaboración.

Adreán Pesántez, quien dio inicio a este estudio, por proporcionarme la matriz de datos morfológicos y su gran colaboración y asesoramiento con sus conocimientos y experiencia.

María Amaguay y todos quienes prestaron su colaboración durante el trabajo de recolección de material en campo.

Al Departamento Nacional de Biotecnología del INIAP en la Estación Experimental "Santa Catalina", en las personas de Eduardo Morillo, por la oportunidad para hacer el análisis molecular, y Johana Buitrón, por las explicaciones sobre marcadores moleculares.

A Miguel Cornejo, por permitirnos el acceso a su finca y donarnos las mazorcas de CCN-51 para la caracterización.

A Víctor Márquez, por las sugerencias iniciales y su buena voluntad y disposición de ayudar.

A Alexander Andrade, por sus buenas recomendaciones y consejos sobre análisis de conglomerados.

A mi madre Yara por todo su amor y sus cuidados, y por mostrarme su cariño y apoyo siempre.
A mis hermanas Carolina y Jennifer, por hacerme sentir su cariño y preocupación.

A mi novio Miguel por ayudarme en lo que necesité y por todo su amor, aliento y comprensión.

A mi tía Jacqueline Obregón por hospedarnos en su departamento durante la recolección de datos.

A todos los docentes que me transmitieron sus conocimientos a lo largo de la carrera, lo cual es invaluable para mí.

Gabriela

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ANEXOS

ÍNDICE DE APÉNDICES

RESUMEN

Se determinaron similitudes entre 26 clones de cacao tipo Criollo y 4 clones testigo referenciales de otros grupos genéticos, con respecto al análisis simultáneo de variables de caracterización morfológica y molecular. Los datos se recolectaron para 38 variables morfológicas medidas sobre mazorcas, semillas, hojas y flores, y para 20 marcadores moleculares microsatélites. De cada matriz morfológica se obtuvo una matriz de distancias poblacionales entre clones, y de la matriz molecular se obtuvo una matriz de similaridad. De cada matriz de distancia/similaridad se obtuvieron las coordenadas principales, siendo cada matriz de coordenadas principales una configuración. Partiendo de éstas, con Análisis Procrustes Generalizado, técnica que permite consensuar estos datos, se obtuvo un consenso entre configuraciones morfológicas, cuyo resultado se comparó con otra metodología aplicada sobre los mismos datos, y finalmente un consenso entre la configuración molecular con el consenso de las configuraciones morfológicas (Software: Excel, R, Infostat). Principales resultados: los mayores coeficientes de similitud molecular se dieron entre: EEB-16 y EEB-21 (0.90), EEB-4 y EEB-10 (0.88), EEB-19 y EEB-25 (0.75), y EEB-2 y EEB-12 (0.73); se hizo un seguimiento a los 6 clones tipo Criollo de mayor rendimiento; el consenso morfológico fue de 83.6%, y el consenso final de 94.3%; tanto en la configuración molecular como en el consenso final EEB-23 no se asemejó a los clones tipo Criollo, estando más cercano al testigo representante del grupo genético Forastero y a CCN-51; los demás clones tipo Criollo conformaron un grupo separado del Forastero, y en general se asemejaron más a los testigos de los grupos genéticos Nacional y Trinitario. La técnica fue apropiada para procesar el tipo de datos obtenidos, por lo que se recomienda utilizarla en estudios similares o en otros donde se necesite analizar diferentes tipos de variables sobre los mismos individuos, simultáneamente.

Palabras clave: <ESTADÍSTICA>, <ANÁLISIS PROCRUSTES GENERALIZADO (APG)>, <ANÁLISIS MULTIVARIANTE>, <CACAO (*Theobroma Cacao L.*)>, <CARACTERIZACIÓN MORFOLÓGICA>, <MARCADORES MOLECULARES MICROSATÉLITES>

SUMMARY

Similarities were determined between 26 Criollo-type cacao clones and 4 referential control clones from other genetic groups, with respect to the simultaneous analysis of morphological and molecular characterization variables. Data were collected for 38 morphological variables measured on pods, seeds, leaves and flowers, and for 20 microsatellite molecular markers. From each morphological matrix, a matrix of population distances between clones was obtained, and a similarity matrix was obtained from the molecular matrix. From each distance/similarity matrix the principal coordinates were obtained, each matrix of principal coordinates being a configuration. Based on these, with Generalized Procrustes Analysis, a technique that allows reach consensus on these data, a consensus was obtained between morphological configurations, whose result was compared with another methodology applied on the same data, and finally a consensus between the molecular configuration and the consensus of the morphological configurations (Software: Excel, R, Infostat). Main results: the highest coefficients of molecular similarity were found between: EEB-16 and EEB-21 (0.90), EEB-4 and EEB-10 (0.88), EEB-19 and EEB-25 (0.75), and EEB-2 and EEB-12 (0.73); the 6 highest yield Criollo-type clones were followed up; the morphological consensus was 83.6%, and the final consensus was 94.3%; both, in the molecular configuration and in the final consensus, EEB-23 did not resemble Criollo type clones, being closer to the control representative of the Forastero genetic group and to CCN-51; the other Criollo-type clones formed a separate group from the Forastero, and in general they resembled more the controls of the Nacional and Trinitario genetic groups. The technique was appropriate to process the type of data obtained, so it is recommended to use it in similar studies or in others where it is necessary to analyze different types of variables on the same individuals, simultaneously.

Keywords: <STATISTICS> <GENERALIZED PROCRUSTES ANALYSIS (GPA)>, <MULTIVARIATE ANALYSIS>, <CACAO (Theobroma Cacao L.)>, MORPHOLOGICAL CHARACTERIZATION ^ <MICROSATELLITE MOLECULAR MARKERS>

INTRODUCCIÓN

Antecedentes del problema

El cultivo de cacao, por su importancia económica, ha merecido la atención del mundo de la ciencia para profundizar cada vez más el conocimiento existente en todos sus aspectos, lo que ha permitido identificar y seleccionar los árboles más idóneos y realizar cruzamientos controlados que se ajusten a los objetivos de producción del agricultor. Estos objetivos están orientados a que los materiales tengan "adecuada adaptación a las diferentes zonas agroecológicas, alta producción, resistencia a las principales enfermedades y buenos atributos de calidad" (QUIROZ, MESTANZA y PARADA 2014).

Si, por ejemplo, se identifica un árbol con atributos de calidad, pero que no resiste a enfermedades y es de baja producción, y por otro lado se tiene un árbol sin atributos de calidad, pero sí resistente y altamente productivo, se puede hacer una hibridación o cruzamiento entre ellos para transferir las cualidades deseadas a un nuevo individuo. Este método tradicional tarda años hasta que crezca la planta para poder evaluar los resultados del cruce, y de cada cruce se obtiene un sinnúmero de plantas.

Con este propósito, el Instituto Nacional Autónomo de Investigaciones Agropecuarias (INIAP) creó el Programa Nacional de Cacao en 1940, y desde entonces se han hecho y se continúan haciendo múltiples estudios para evaluar, seleccionar, mejorar y conservar materiales que presentan las características deseadas y poder recomendarlos para su cultivo.

En uno de los estudios recientes, realizado por el Programa de Cacao en la Estación Experimental Litoral Sur "Dr. Enrique Ampuero Pareja" del INIAP, en 2009 se plantó un jardín clonal con 26 clones de cacao fino de aroma de tipo Criollo provenientes de la colección de germoplasma de Esmeraldas, con el fin de realizar una pre-selección de los clones con características deseadas como buen rendimiento y resistencia a enfermedades, a través de la determinación de diferencias y similitudes fitosanitarias, agronómicas, morfológicas y de rendimiento (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014).

Pesántez[1] indicó que en este jardín ya se estudiaron los aspectos: agronómico (vigor del tallo, altura de planta, floración, emisión de follaje, etc.), productivo, fitosanitario (resistencia a monilla, fitóftora, escoba de bruja), para seleccionar los clones más productivos. De estos estudios

[1] En una entrevista personal.

se pudo determinar clones resistentes a *Moniliophthora roreri*, o que los clones EEB-4, EEB-16 y EEB-21 presentan buenas características organolépticas para comercio (frutales y florales).

Además de esto Pesántez (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014) desarrolló el trabajo de tesis sobre estos materiales y 6 clones testigo sobre los cuales realizó una caracterización morfológica y de rendimiento, donde se determinaron los cinco materiales de mayor rendimiento y nueve grupos mediante agrupamiento jerárquico de Ward (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014).

Con base en lo anterior en el 2015 se hicieron hibridaciones o cruces controlados entre Criollos y Nacionales altamente productivos y recomendados para diferentes zonas. En consecuencia, todo esto representa el avance que se tiene hasta el momento con respecto a este jardín clonal de cacao tipo Criollo.

Sin embargo, para aportar con mayor información para este estudio, además de la caracterización morfológica, que está influenciada por el medio ambiente, se requiere complementarla con la caracterización molecular, que no se había realizado antes.

Respecto a la caracterización morfológica se sabe que existen variables como la forma de la mazorca, la pigmentación de la flor o el color de la almendra, que no están muy influenciadas por el medio ambiente. Por el contrario, variables como largo de mazorca o número de semillas por mazorca sí pueden estarlo debido a factores como un mejor suelo o mayor luminosidad.

Por su parte, el enfoque de la caracterización con marcadores moleculares puede contribuir a determinar qué tanto se asemejan los clones a los grandes grupos genéticos que se conocen: Criollo, Trinitario, Forastero y Nacional, pero a nivel de ADN, el cual, a diferencia de la parte morfológica, no es influido por el medio ambiente.

Considerando estos aspectos, en el International Plant Genetic Resources Institute (IPGRI) (INTERNATIONAL PLANT GENETIC RESOURCES INSTITUTE (IPGRI) 2003) se establece que analizar estadísticamente datos de caracterización morfológica es el primer paso para estudiar la variabilidad genética, pero para un mejor entendimiento de ésta es igualmente importante analizar datos de caracterización molecular, pues un análisis integrado que incluya varias disciplinas permite una mejor comprensión de la variabilidad genética. En este sentido, Pesántez (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014) luego de haber realizado la caracterización morfológica de este jardín clonal, recomienda complementarla con la caracterización molecular para determinar la correlación.

Son muy necesarias herramientas estadísticas que permitan analizar la mezcla de información agronómica y molecular que se genera con la utilización de ambos descriptores para lograr una descripción más completa, existiendo la necesidad de tratar estos datos de forma articulada (BRAMARDI, y otros 2005). El problema que surge al tratar conjuntamente información de diferente naturaleza es que para variables cuantitativas es más apropiado trabajar con distancias, para variables cualitativas es más apropiado trabajar con similaridades, y para variables moleculares lo más apropiado es utilizar similaridades genéticas; además si de cada tipo de variable se tienen diferentes cantidades de variables, hay métodos que se pueden ver afectados al dar más peso al que más variables tenga (que no necesariamente significa que proporcione mayor información genética). Según Bramardi et al. (BRAMARDI, y otros 2005) una técnica que supera estas dificultades es el Análisis Procrustes Generalizado (APG) al permitir llegar a un consenso entre la información agronómica y molecular, pues permite utilizar las diferentes distancias/similaridades que son más apropiadas para cada tipo de variable, y no da una mayor ponderación al tipo de información que más variables tenga. Dichos autores evaluaron el APG, utilizando datos de variedades conocidas de cultivos de pepinillo, al compararlo con las técnicas: extensión del Coeficiente de Emparejamiento Simple, y Análisis de Correspondencias, por medio de la correlación entre matrices de distancias/similaridades con significancia del test de Mantel. Se concluyó que al caracterizar individualmente la información agronómica y molecular los resultados fueron diferentes y no se calsificaron apropiadamente los cultivos, pero caracterizándolos simultáneamente con APG la clasificación fue más efectiva, y además superó las desventajas mencionadas que tenían las otras técnicas (BRAMARDI, y otros 2005). El APG al ser una técnica descriptiva no requiere del cumplimieno de supuestos.

De igual manera Bruno y Balzarini (BRUNO y BALZARINI 2010) recomiendan el empleo de APG para "ordenar en un único espacio las observaciones según la información brindada tanto por marcadores morfológicos como moleculares", afirman que "la interpretación de la información provista por múltiples marcadores mejora sustancialmente al poder visualizar en un espacio de baja dimensión las observaciones" y concluyen que "el APG es una técnica útil cuando se desean estudiar las relaciones entre materiales a partir de datos de marcadores de diferente naturaleza" (BRUNO y BALZARINI 2010).

Lo que hace el APG de forma general es consensuar dos o más matrices que contienen datos sobre los mismos individuos, pero cuyas valores observados difieren de una matriz a otra, lo cual se puede deber a que las variables en cada matriz son de diferente naturaleza, reflejan diferentes aspectos de los individuos, o podría tratarse de las mismas variables pero medidas con diferentes instrumentos de medición en cada matriz (como en análisis de perfil sensorial donde cada juez se puede ver como el instrumento de medición). No importa si el número de variables no es el mismo

en cada matriz. Además, la o las matrices pueden partir de variables cualitativas si se obtienen las coordenadas principales de su respectiva matriz de distancia o similaridad.

Bramardi, et al. (BRAMARDI, y otros 2005) afirman que hasta ese año el APG era muy utilizado en otros campos de aplicación, principalmente en análisis de perfil sensorial, en geometría morfométrica y en alineación molecular en estudios de estructura-actividad, y de lo que se conocía, el de dichos autores fue uno de los primeros trabajos donde se aplicó en agronomía con información morfológica y molecular. Se realizó una revisión sistemática hasta el año 2017 donde se encontró que, además de en los campos ya mencionados, se ha aplicado en análisis de forma y análisis de imágenes; además se encontró que la gran mayoría de publicaciones se tratan de una aplicación de la técnica, y muy pocas de desarrollo teórico de la misma. De lo último se destaca la publicación de Xiong et al. (XIONG, y otros 2008) donde se propuso un test de permutación que permite probar hipótesis nulas específicas dentro de APG; las demás publicaciones adaptan la técnica haciéndola más específica para cada campo de aplicación.

Hasta la realización de este trabajo se encontró que existen varios estudios donde se aplicó esta técnica con información morfo-agronómica y molecular, desde hace algunos años en cultivos de: yuca (DEMEY, y otros 2003), pepinillo (BRAMARDI, y otros 2005), cacao boliviano (JULY MARTINEZ 2007), quinua (COSTA TARTARA, y otros 2011), alfalfa (GRANDON, y otros 2013), tomate (MAHUAD, y otros 2013) y cactus (JANA, y otros 2017).

Justificación de la investigación

La importancia estadística de este trabajo radica en la metodología aplicada y en la forma de tratar y analizar los datos disponibles mediante las técnicas estadísticas más apropiadas según la naturaleza y estructura de los mismos para un buen aprovechamiento de la información. Con respecto a los datos morfológicos que ya fueron analizados previamente por otro autor, se realiza una comparación de los resultados obtenidos utilizando diferentes técnicas y dando un tratamiento distinto al mismo conjunto de datos de partida. Se comparan los resultados previos con los resultados aquí obtenidos mediante el empleo de escalado multidimensional con distancia entre dos poblaciones (Mahalanobis), para aplicar Análisis Procrustes Generalizado sobre las matrices de coordenadas principales, y lograr un consenso, con base en el cual se pueda realizar una mejor clasificación de los árboles clonados de cacao, pues lo que interesa agronómicamente hablando es que para esto se consideren simultáneamente todas las variables de importancia agronómica. Siendo el aporte adicional de este trabajo la información proporcionada por marcadores moleculares, se busca también la forma más apropiada para tratar este tipo de datos según lo recomendado en la literatura, para con ello lograr un consenso final. Además de esto, con enfoque

en las variables morfológicas, se realizan test no paramétricos alternativos a los empleados por el anterior autor, para de igual manera comparar los resultados obtenidos.

Con respecto al cacao, cuando éste es de calidad se lo conoce como *fino y de aroma* si presenta "características individuales distintivas, de toques florales, frutales, nueces, almendras y especias" (ASOCIACIÓN NACIONAL DE EXPORTADORES DE CACAO - ECUADOR (ANECACAO) s.f.a), lo cual es muy apetecido por los grandes chocolateros del mundo para la elaboración de chocolates finos (EL CIUDADANO 2015), altamente cotizado en Estados Unidos y Europa y por el que el mercado gourmet paga sobreprecios (VAZQUEZ OVANDO, y otros 2012). Al cacao que no presenta características de calidad se lo conoce como *al granel* o *básico*. Se estima que el 95% de la producción mundial corresponde a cacao al granel, y sólo el 5% a cacao fino y de aroma (LOZADA VARGAS 2014) aunque según CasaLuker (CASALUKER s.f.) representa alrededor del 8%.

Ecuador es conocido internacionalmente como uno de los mayores productores de cacao fino y de aroma del mundo, pues aporta con más del 70% de la producción mundial del mismo (ASOCIACION NACIONAL DE EXPORTADORES DE CACAO - ECUADOR (ANECACAO) s.f.b) cuando el 76% del cacao fino de aroma del mundo lo producen entre Ecuador, Colombia, Venezuela y Perú (CASALUKER s.f.). En 2015 desde Ecuador se exportó un volumen total de 260 mil toneladas métricas de cacao en grano y productos derivados de cacao (ASOCIACION NACIONAL DE EXPORTADORES DE CACAO - ECUADOR (ANECACAO) s.f.b).

Respecto a los cuatro grandes grupos genéticos de cacao, el grupo Forastero, caracterizado por ser altamente productivo, es considerado cacao al granel, mientras que los grupos Nacional, Criollo y Trinitario (que desciende de un cruce entre Criollo y Forastero) son considerados finos y de aroma. Por ello la importancia de estudiar los 26 materiales de cacao tipo Criollo. Se estima que existen 244 hectáreas de cacao Criollo en el país (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014).

El mejor cacao para hacer chocolate es el Criollo, pero una desventaja es que es muy susceptible a enfermedades (*Moniliophtora roreri* y *Phytophtora palmivora*) y es de bajo vigor, lo cual es muy limitante porque tiene menor rendimiento agronómico (mazorcas más pequeñas y menor cantidad de almendras por fruto) (VAZQUEZ OVANDO, y otros 2012). Esto se corrobora en el estudio de rendimiento de los 26 materiales de tipo Criollo hecho por Pesántez (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014) como lo muestra el **Gráfico 1**.

Por eso, con base en la caracterización, se quieren hacer cruzamientos con otros materiales que no tienen esta desventaja para mejorar las características de los cultivos, lo que se conoce como mejoramiento de plantas o fitomejoramiento.

Caracterizar las colecciones de germoplasma es un paso fundamental en el manejo de colecciones, pues al conocerlas morfológicamente se puede "depurar u organizar los materiales y sobre todo identificar genotipos valiosos para ser usados directamente o utilizarlos en programas de mejoramiento genético" por lo que "es vital tener información disponible de cada material, sobre caracteres cualitativos y cuantitativos de importancia actual o futura." (GUACHO ABARCA 2014)

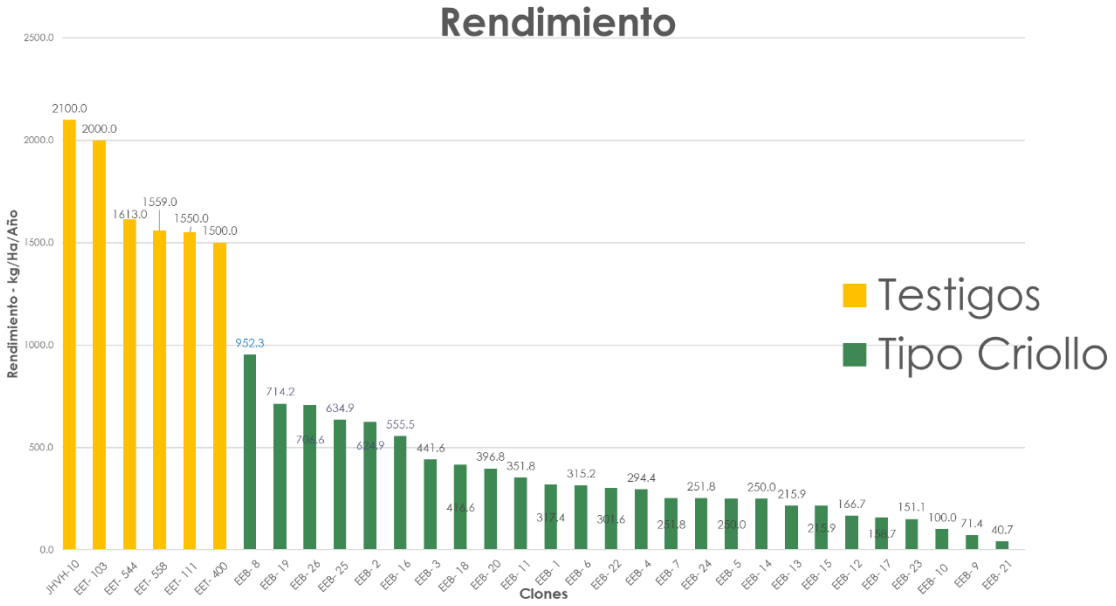

Gráfico 1: Comparación del rendimiento de los 26 materiales tipo Criollo con varios testigos.

Realizado por: Gabriela J. Obregón O. 2018

Este trabajo presta su utilidad al generar información sobre las similitudes entre clones con respecto a las caracterizaciones morfológica y molecular dando un tratamiento a los datos con Análisis Procrustes Generalizado, técnica apropiada para tratar conjuntos de variables de diferente naturaleza medidas sobre las mismas unidades estadísticas, llegando a un consenso entre los conjuntos de variables, con lo que se pueden determinar similitudes más reales al considerar simultáneamente datos morfológicos y datos de ADN. Respecto a la aplicación de esta técnica en cultivos solo se la encontró en los mencionados previamente (yuca, pepinillo, cacao boliviano, quinua alfalfa, tomate y cactus) pero no se encontró una aplicación de esta técnica en cacao de Ecuador.

La información resultante sirve de apoyo a los fitomejoradores durante el proceso de selección de plantas con características deseadas para decidir cuáles serán los padres que transmitirán las características deseadas a los nuevos individuos al obtener distancias más reales entre clones por considerar simultáneamente diferentes tipos de descriptores. Además de esto, el aporte de la

caracterización molecular hace posible distinguir individuos a nivel de ADN y en un tiempo muy corto en comparación a lo que tarda la caracterización morfológica.

En el país el cacao fino y de aroma es cultivado por 100 000 familias, de las cuales el 99% son pequeños productores (MINISTERIO DE AGRICULTURA Y GANADERÍA (MAGAP) 2013), así lo que se espera es que este aporte contribuya en última instancia a beneficiar a los pequeños agricultores, que son los guardianes de la variabilidad genética que es tan importante para la conservación de un cultivo, al aportar con la caracterización que es parte del proceso de selección y mejoramiento que pretende lograr materiales con características idóneas para ellos.

Identificación del problema

Hoy en día es muy difícil encontrar materiales puros debido a la gran mezcla por los cruces que se han dado naturalmente. Por eso la variación genética es grande, y en Ecuador hay una gran colección. De ahí surge la necesidad de distinguir los 26 materiales comparándolos con los tipos puros que ya están muy bien identificados, para saber si se asemejan más a Criollo, como se supone a priori, o si por el contrario uno o más presentan mayor similitud con Nacional, Trinitario o Forastero.

Si se tiene información morfológica y molecular de estos 26 materiales de tipo[2] Criollo, es necesario emplear una técnica que permita hacer el análisis tomando en cuenta ambos tipos de datos, que son de diferente naturaleza, para poder concluir qué tan similares son a Criollo, o ver si tienen mayor similitud con los testigos representativos de cada grupo genético.

Objetivos

Objetivo general

Investigar la similitud de los 26 clones de cacao de tipo Criollo y los 4 clones testigo, considerando simultáneamente marcadores morfológicos y marcadores moleculares.

Objetivos específicos

[2] Se antepone la palabra *tipo* porque no son 100% puros

- Preparar las matrices de datos retirando valores incongruentes y variables redundantes y encontrando las matrices de distancias/similaridades por clon con respecto a cada unidad de análisis, en R y Excel.
- Obtener la solución por coordenadas principales de cada matriz de distancia/similaridad y comparar la similitud entre clones en marcadores moleculares y morfológicos por separado.
- Llegar a un consenso entre los tipos de marcadores con Análisis Procrustes Generalizado (en Infostat) a partir de las soluciones por coordenadas principales para evaluar la similitud entre clones.
- Determinar conglomerados con agrupamiento jerárquico de Ward a partir de las configuraciones individuales y del consenso para comparar los grupos de clones que se forman en cada uno.

Orientación al lector

En el Capítulo I se aborda el marco teórico referencial donde se detalla la teoría estadística y conceptos importantes encontrados en la revisión de la literatura. En el Capítulo II se explica la metodología que se siguió. En el Capítulo III se presentan los resultados obtenidos y discusión. En esta última se comenta y explica la diferencia en el análisis morfológico de este trabajo con el del autor que previamente realizó el mismo utilizando los mismos datos. Finalmente se muestran las conclusiones y recomendaciones. Algunos códigos de R utilizados en el análisis se adjuntan en el apéndice de este documento, y otros en la *nube* en la siguiente dirección: https://onedrive.live.com/?authkey=%21AIJE2OcZQnWrOgE&id=CA77A5953C07ED0C%21140440&cid=CA77A5953C07ED0C.

CAPÍTULO I

1 MARCO TEÓRICO REFERENCIAL

1.1 Análisis Multivariante

Se trata de un conjunto de técnicas que permiten representar y sintetizar múltiples caracteres observados sobre un colectivo de unidades estadísticas. Las técnicas de análisis se pueden clasificar como: descriptivas cuando se trabaja con variables estadísticas multidimensionales o vectores estadísticos y el colectivo es la población; e inferenciales cuando se trabaja con variables aleatorias multidimensionales o vectores aleatorios y el colectivo es la muestra.

En el **Gráfico 1-1** se presentan los tipos de variables, mismos que dependen del instrumento de medida.

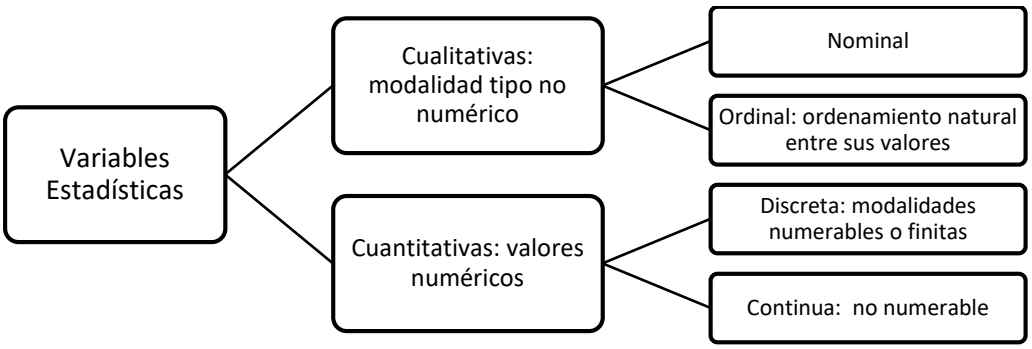

Gráfico 1-1: Tipos de variables estadísticas.

Realizado por: Gabriela J. Obregón O. 2018

Sobre un colectivo de tamaño n se miden p variables. Los datos se organizan en una tabla a partir de la cual se construye una matriz de datos, así se da el paso de la estadística hacia el cálculo matricial (**Gráfico 2-1**). $i = 1, \dots, n$ indica el individuo y $j = 1, \dots, p$ indica la variable.

Individuo	X_1	...	X_p
1	x_{11}	...	x_{1p}
.	.	.	.
.	.	.	.
.	.	.	.
n	x_{n1}	...	x_{np}

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)'$$

Cada vector columna $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ con $i = 1, 2, \dots, n$ es una p-upla de números reales y corresponde al dato i de la distribución multivariada.

Gráfico 2-1: Paso de una tabla estadística hacia una matriz.

1.1.1 Conceptos importantes

1.1.1.1 Variable multidimensional

$\boldsymbol{X} = (X_1, \dots, X_p)$ es una variable estadística p-dimensional, cuyas componentes son variables estadísticas unidimensionales que asocian a cada individuo un número. Así, la variable estadística p-dimensional asocia a cada unidad estadística una p-upla de números.

1.1.1.2 Distribución estadística multivariada

Se denota como $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{n}$.

1.1.1.3 Vector de medias

Dada una $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{n}$ de una variable estadística p-dimensional $\boldsymbol{X} = (X_1, \dots, X_p)$ con componentes cuantitativas, se llama vector de medias al vector $p \times 1$:

$$\bar{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = (\bar{x}_1, \dots, \bar{x}_p)'$$

En términos de la matriz de datos **X**:

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}$$

Donde se encuentra la transpuesta de la matriz de datos

$$\mathbf{X}' = \begin{pmatrix} x_{1\,1} & \cdots & x_{1\,n} \\ \vdots & \ddots & \vdots \\ x_{p\,1} & \cdots & x_{p\,n} \end{pmatrix} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$$

Se multiplica por el vector de unos de orden $n \times 1$:

$$\mathbf{1} = (1, 1, \dots, 1)'$$

Y se multiplica el escalar $\frac{1}{n}$ por el resultado de ese producto.

Es un indicador de posición multivariado que tiene como componentes las medias de la distribución.

1.1.1.4 Matriz de centrado

Es la matriz $\mathbf{H} = \mathbf{I}_{n\times n} - \frac{1}{n}\mathbf{J}$ (Siendo **J** una matriz cuadrada de unos).

1.1.1.5 Matriz de datos centrados de $\boldsymbol{x}$

Se denota con $\widetilde{\mathbf{X}}$ y se calcula con $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$

1.1.1.6 Matriz de covarianzas

Dada una $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{n}$ de una variable estadística p-dimensional $\boldsymbol{X} = (X_1, \dots, X_p)$ con componentes cuantitativas, se llama matriz de covarianzas de **x** a la matriz cuadrada de orden p:

$$\mathbf{S} \coloneqq \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_\mathrm{i} - \bar{\mathbf{x}})\,(\mathbf{x}_\mathrm{i} - \bar{\mathbf{x}})' = \frac{1}{n}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \begin{pmatrix} S_{11} & \cdots & S_{1p} \\ \vdots & \ddots & \vdots \\ S_{p1} & \cdots & S_{pp} \end{pmatrix} = \begin{pmatrix} \mathrm{S}_1^2 & \cdots & S_{1p} \\ \vdots & \ddots & \vdots \\ S_{p1} & \cdots & \mathrm{S}_p^2 \end{pmatrix}$$

Sus propiedades son: es una matriz cuadrada de orden p, es simétrica (coincide con su transpuesta), contiene en su diagonal principal las varianzas de las variables, fuera de la diagonal contiene las covarianzas entre los pares de variables, es semi-definida positiva: $\forall\, \mathbf{y} \in \mathbb{R}^p$ se tiene $\mathbf{y}'\,\mathbf{S}\mathbf{y} \geq 0$, su traza es no negativa: $tr(\mathbf{S}) \geq 0$, su determinante es no negativo: $det(\mathbf{S}) \geq 0$, sus autovalores son no negativos.

Es la matriz más importante del Análisis Multivariante por contener toda la información de variabilidad.

1.1.1.7 Matriz de correlación

Es una matriz cuadrada de orden p, es simétrica, tiene unos en la diagonal principal, fuera de la diagonal contiene los coeficientes de correlación lineal entre los pares de variables.

$$\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}} = \begin{pmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & r_{pp} \end{pmatrix} = \begin{pmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{pmatrix}$$

Donde: $\mathbf{D}^{-\frac{1}{2}} = diag(S_1^{-1}, S_2^{-1}, \dots, S_p^{-1})$, es decir, contiene las inversas de las desviaciones típicas.

Teorema de la dimensión

Si $r = Rng(\mathbf{S}) \leq p$ significa que hay r variables linealmente independientes, y las otras $p - r$ son combinación lineal de estas variables independientes.

Se puede determinar lo dicho por el teorema de la dimensión según el número de autovalores aproximados a cero, y los autovectores asociados a los autovalores nulos proporcionan los coeficientes de las combinaciones lineales.

1.1.2 Medidas globales de variabilidad y dependencia

1.1.2.1 Varianza generalizada

Es el determinante de la matriz de covarianzas **S**: $|\mathbf{S}|$

1.1.2.2 Varianza total

Es la traza de la matriz de covarianzas **S**: $tr(\mathbf{S})$

1.1.2.3 Coeficiente de dependencia

$\eta^2 = 1 - |\mathbf{R}|$, donde $|\mathbf{R}|$ es el determinante de la matriz de correlación **R**. Propiedades: $0 \leq \eta^2 \leq 1$, $\eta^2 = 0$ si y sólo si las p variables están incorreladas, $\eta^2 = 1$ si y sólo si hay relaciones lineales entre las variables.

1.1.3 Estandarización Univariada

De forma multivariada se considera la siguiente transformación lineal:

$$\mathrm{Y} = \mathbf{D}^{-1/2}(\mathrm{X} - \bar{\mathbf{x}})$$

Donde $\mathbf{D}^{-1/2} = diag(S_1^{-1}, S_2^{-1}, \dots, S_p^{-1})$, es decir, una matriz diagonal que contiene las inversas de las desviaciones típicas de las variables $X_1, \dots, X_p$.

La nueva matriz de datos es igual a $\mathbf{Y} = \tilde{\mathbf{X}}\mathbf{D}^{-1/2}$. El nuevo vector de medias es $\bar{\mathbf{y}} = \mathbf{D}^{-\frac{1}{2}}(\bar{\mathbf{x}} - \bar{\mathbf{x}}) = \mathbf{0}$, es decir, un vector de ceros. Y la nueva matriz de covarianzas es $\mathbf{S_Y} = \mathbf{R_X} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$, es decir, es la matriz de correlación de la matriz de datos original. En conclusión, lo que se logra

es que todas las variables tengan media 0 y varianza 1. No toma en cuenta las relaciones de dependencia lineal.

1.2 Análisis de Componentes Principales

El propósito es, siendo p el número de variables en estudio, construir nuevas variables (llamadas Componentes Principales) en número menor que p, de tal manera que se cumplan dos condiciones fundamentales:

1. Que permitan explicar la mayor parte de la variabilidad de los datos minimizando la pérdida de información.
2. Que permitan estudiar la distribución multivariada representándola en un espacio de dimensión más pequeño que p, y que esta representación sea lo más ajustada posible a la original, para lo cual se garantiza que las distancias entre los elementos de la transformación sean aproximadamente iguales a como lo eran originalmente (antes de la transformación). Por ejemplo, si $p = 10$, que mediante 2 nuevas variables o en 2 dimensiones los elementos se puedan representar óptimamente.

Estas nuevas variables se obtienen como combinaciones lineales de las originales.

Las COMPONENTES PRINCIPALES (CP) son las variables $Y_1 = \mathbf{t}_1'\mathrm{X}$, $Y_2 = \mathbf{t}_2'\mathrm{X}, \dots, Y_p = \mathbf{t}_p'\mathrm{X}$ tales que:

1. $Var(Y_1)$ es máxima condicionada a $\mathbf{t}_1'\mathbf{t}_1 = 1$, es decir tiene máxima varianza y $\mathbf{t}_1$ tiene norma unidad.
2. Entre todas las variables compuestas Y tales que $Cov(Y_1, Y) = 0$, la variable Y_2 es tal que $Var(Y_2)$ es máxima condicionada a $\mathbf{t}_2'\mathbf{t}_2 = 1$, es decir que Y_2 está incorrelada con Y_1 y debe tener varianza máxima, y $\mathbf{t}_2$ tiene norma unidad.
3. Y_3 es una variable incorrelada con Y_1, Y_2, con máxima varianza y $\mathbf{t}_3$ con norma unidad, y así, análogamente, se definen las demás Componentes Principales.

Así se tiene un vector de CP: $\mathrm{Y} = (Y_1, \dots, Y_p)'$. Para la transformación: Sea $\mathbf{T} = \left(\mathbf{t}_1, \dots, \mathbf{t}_p\right)$ la matriz $p \times p$ cuyas columnas son los vectores que definen las componentes principales, la transformación de X (vector de variables), $\mathrm{Y} = \mathbf{T}\mathrm{X}$ se llama TRANSFORMACIÓN POR CP.

El teorema nos dice cómo deben ser los vectores $\mathbf{t}_1, \dots, \mathbf{t}_p$, componentes de $\mathbf{T}$.

Teorema

Sean $\mathbf{t}_1, \ldots, \mathbf{t}_p$ los p autovectores normalizados (con norma 1) de la matriz de covarianzas **S**, es decir, $\mathbf{S}\mathbf{t}_i = \lambda_i \mathbf{t}_i$ y $\mathbf{t}_i'\mathbf{t}_i = 1$ con $i = 1, \ldots, p$. Entonces:

1. Las variables compuestas $Y_i = \mathbf{t}_i'\mathrm{X}$ con $i = 1, \ldots, p$ son los CP.
2. Las varianzas de éstas son los autovalores de **S**. $Var(Y_i) = \lambda_i$ con $i = 1, \ldots, p$.
3. Las CP son variables incorreladas: $Cov(Y_i, Y_j) = 0$ con $i \neq j = 1, \ldots, p$.

Ahora se sabe que los componentes de la matriz **T** son los autovectores de la matriz de covarianzas $\mathbf{S_X}$, y que la matriz de covarianzas de las nuevas variables (CP) es una matriz diagonal que contiene los autovalores de $\mathbf{S_X}$.

Matricialmente las transformaciones se expresan como: $\mathbf{Y} = \mathbf{X}(\mathbf{T}')' = \mathbf{XT}$; $\bar{\mathbf{y}} = \mathbf{T}'\bar{\mathbf{x}}$; $\mathbf{S_Y} = \mathbf{T}'\mathbf{S_X}\mathbf{T}$.

Para garantizar la primera condición, se tiene que:

Es de interés saber qué cantidad de variabilidad está siendo explicada. La variabilidad total explicada es la traza de la matriz de covarianzas: $tr(\mathbf{S_X}) = \sum_{i=1}^{p} \lambda_i$. La variabilidad explicada por la $i - ésima$ CP es $Var(Y_i) = \lambda_i$. Así la fracción de variabilidad total explicada por la CP Y_i viene dada como $\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$. Si denotamos a las CP como $Y_1, \ldots, Y_q, Y_{q+1}, \ldots, Y_p$, la fracción de variabilidad total explicada por las q primeras CP es $\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$.

Si con las 3 primeras CP no se alcanza a representar una fracción alta de la variabilidad total, hay que recurrir a CP no lineales.

Para garantizar la segunda condición nos basamos en la distancia euclídea, y se tiene que:

Sea la distancia euclídea al cuadrado entre la fila i $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ y la fila j $\mathbf{x}_j' = (x_{j1}, \ldots, x_{jp})$ de la matriz de datos **X**:

$$\delta_{ij}^2 = \left(\mathbf{x}_i - \mathbf{x}_j\right)'\left(\mathbf{x}_i - \mathbf{x}_j\right) = \sum_{h=1}^{p} (x_{ih} - x_{jh})^2$$

1.2.1 Matriz de distancias entre filas

Se define a la MATRIZ DE DISTANCIAS ENTRE FILAS de dimensiones $n \times n$ como: $\mathbf{\Delta} = (\delta_{ij})$, es decir, con componentes raíz cuadrada de lo anterior. Con base en esto se define la variabilidad geométrica de una matriz de distancias.

VARIABILIDAD GEOMÉTRICA: La variabilidad geométrica de una matriz de distancias **Δ** es $V_\delta(\mathbf{X}) = \frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij}^2$.

El teorema permite calcular la variabilidad geométrica.

Teorema

La variabilidad geométrica de la matriz de distancias euclídeas Δ es la traza de la matriz de covarianzas:

$$V_\delta(\mathbf{X}) = tr(\mathbf{S}) = \sum_{\mathrm{i}=1}^{\mathrm{p}} \lambda_i$$

Si la representación tiene máxima variabilidad geométrica se logra el objetivo de que las distancias en dimensión reducida sean iguales a las distancias originales, y resulta que el siguiente teorema nos dice que la transformación por CP es la que maximiza la variabilidad:

Teorema

La transformación lineal **T** que maximiza la variabilidad geométrica en dimensión q con $q \leq p$ es la transformación por CP, es decir $\mathbf{T} = (\mathbf{t}_1, \dots, \mathbf{t}_q)$ contiene los q primeros autovectores normalizados de **S**.

Esto garantiza una representación óptima en dimensión q ($q \leq p$). La fracción de variabilidad geométrica explicada por la nueva matriz de datos **Y** es: $P_q = \frac{V_\delta(\mathbf{Y})_q}{V_\delta(\mathbf{X})}$

Las CP, además de conservar la variabilidad original de los datos $tr(\mathbf{S_Y}) = tr(\mathbf{S_X})$, también conservan la varianza generalizada $|\mathbf{S_Y}| = |\mathbf{S_X}|$.

En caso de haber muchas diferencias entre las intensidades totales de las variables o entre sus varianzas, conviene hacer un ANÁLISIS NORMADO, lo cual se logra haciendo una estandarización univariante antes de aplicar el ACP.

1.3 Noción general de distancia

Sea X un conjunto no vacío. Una distancia (o métrica) sobre X es una función $d: X \times X \to \mathbb{R}$ que satisface las siguientes condiciones:

1. $d(x, y) \geq 0 \; \forall \; x, y \in X$;
2. $d(x, y) = d(y, x)$;

3. $d(x,y) \leq d(x,z) + d(z,y)\ \forall\ x,y,z \in X$ (desigualdad triangular);
4. $d(x,y) = 0 \Leftrightarrow x = y$

Así el número real no negativo $d(x,y)$ se llama distancia entre x y y.

En el contexto del Análisis Multivariante se tienen los datos $\mathbf{x}_i' = (x_{i1}, \dots, xi_p)$ y $\mathbf{x}_j' = (x_{j1}, \dots, x_{jp})$. Se definen las siguientes distancias:

1.3.1 Distancia Euclídea

$$d_E(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2} = \sqrt{\sum_{h=1}^{p} (x_{ih} - x_{jh})^2}$$

1.3.2 Distancia de Mahalanobis

$$d_M(i,j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$

La primera supone que las variables están incorreladas, por lo que sólo es útil en tal caso. La segunda tiene una ventaja sobre la otra porque considera las correlaciones, o sea, la covarianza.

1.3.3 Distancia entre dos poblaciones

Suponiendo dos poblaciones representadas por dos matrices de datos: $\mathbf{X}$ de orden $n_1 \times p$ y $\mathbf{Y}$ de orden $n_2 \times p$. La distancia entre las dos poblaciones queda definida por: $(\bar{\mathbf{x}} - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})$ donde $\bar{\mathbf{x}}$ y $\bar{\mathbf{y}}$ son los vectores de medias, y $\mathbf{S} = \frac{n_1 \mathbf{S}_1 + n_2 \mathbf{S}_2}{n_1 + n_2}$ es una media ponderada de las matrices de covarianzas de las distribuciones.

1.4 Escalado multidimensional

El objetivo es el mismo que en ACP: encontrar una representación de los individuos en un espacio de dimensión reducida de tal manera que las distancias sean muy parecidas a las originales. A diferencia del ACP donde se dispone de una matriz de orden $n \times p$ de variables sobre un colectivo, en el escalado multidimensional se dispone de una matriz de distancias entre

individuos. Estas distancias podrían ser, por ejemplo, distancias euclídeas, pero también pueden ser disimilaridades. Se hace necesario identificar cuándo una función es una distancia. Recordando la noción general de distancia: $d: X \times X \to \mathbb{R}$ como una función que satisface las condiciones:

1. $d(x,y) \geq 0 \; \forall \; x,y \in X$;
2. $d(x,y) = d(y,x)$;
3. $d(x,y) \leq d(x,z) + d(z,y) \; \forall \; x,y,z \in X$ (desigualdad triangular);
4. $d(x,y) = 0 \Leftrightarrow x = y$

Una distancia cumple con las cuatro condiciones (Euclídea, Mahalanobis).

Se considera, entonces, la MATRIZ SIMÉTRICA DE DISTANCIAS (o disimilaridades):

$$\boldsymbol{\Delta} = \begin{pmatrix} \delta_{11} & \cdots & \delta_{1n} \\ \vdots & \ddots & \vdots \\ \delta_{n1} & \cdots & \delta_{nn} \end{pmatrix}$$

Y de $\boldsymbol{\Delta}$ se dice que es una matriz de distancias euclídeas si existen n puntos $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ siendo $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ con $i = 1, \ldots, n$, tales que

$$\delta_{ij}^2 = \sum_{\alpha=1}^{p} (x_{i\alpha} - x_{j\alpha})^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$$

Así, lo que se busca es una matriz de datos, la mejor matriz que haga que las distancias euclídeas sean lo más similares a las distancias dadas entre individuos. Se verá a continuación cuándo o bajo qué condiciones podríamos afirmar que $\boldsymbol{\Delta}$ es de distancias euclídeas.

Si $\boldsymbol{\Delta} = (\delta_{ij})$, $\boldsymbol{\Delta}^{(2)} = (\delta_{ij}{}^2)$. Bajo el supuesto de que $\boldsymbol{\Delta}$ sí sea de distancias euclídeas, de acuerdo a lo establecido cada elemento de $\boldsymbol{\Delta}^{(2)}$ se obtiene $\delta_{ij}{}^2 = \mathbf{x}_i'\mathbf{x}_i + \mathbf{x}_j'\mathbf{x}_j - 2\mathbf{x}_i'\mathbf{x}_j$. La matriz de productos escalares asociada a $\boldsymbol{\Delta}$ es $\mathbf{G} = \mathbf{X}\mathbf{X}'$ (donde se está suponiendo que $\mathbf{X} = \begin{pmatrix} x_{1\,1} & \cdots & x_{1\,p} \\ \vdots & \ddots & \vdots \\ x_{n\,1} & \cdots & x_{n\,p} \end{pmatrix}$). $\mathbf{G} = (g_{ij})$ es de productos escalares porque sus elementos son productos escalares entre los puntos, es decir, $g_{ij} = \mathbf{x}_i'\mathbf{x}_j$. Ahora interesa encontrar una relación entre esta $\mathbf{G}$ y esta $\boldsymbol{\Delta}$. Consideremos el vector $\mathbf{g}$ formado por los elementos $\mathbf{g} = (g_{11}, \ldots, g_{nn})$. La relación que existe es: $\boldsymbol{\Delta}^{(2)} = \mathbf{1}\mathbf{g}' + \mathbf{g}\mathbf{1}' - 2\mathbf{G}$ (siempre bajo la suposición de que $\boldsymbol{\Delta}$ es de distancias euclídeas).

Considerar las siguientes matrices:

$$\mathbf{A} = -\frac{1}{2}\boldsymbol{\Delta}^{(2)}$$

$$\mathbf{B} = \mathbf{HAH}$$

Donde **H** es la matriz de centrado $\mathbf{H} = \mathbf{I}_{n\times n} - \frac{1}{n}\mathbf{J}$. Se enuncia el teorema.

Teorema

La matriz **Δ** es una matriz de distancias euclídeas si y sólo si $\mathbf{B} \geq 0$, es decir, si es semidefinida positiva.

Con base en esto lo que se hace es construir **A**, luego construir **B** y con eso, si **B** es semidefinida positiva, entonces **Δ** es de distancias euclídeas, caso contrario, no existen los puntos $\mathbf{x}_\mathrm{i}$, $\mathbf{x}_\mathrm{j}$, etc. Si **B** es semidefinida positiva, sus autovalores son positivos.

Sea $\mathbf{B} \geq 0$ de rango p, entonces ∃ una matriz $n \times p$ llamada **Y**, de rango p, tal que **B** se puede escribir como $\mathbf{B} = \mathbf{Y}\mathbf{Y}'$. Ahora bien, consideremos la descomposición espectral (que vale para cualquier matriz simétrica) de **B**: $\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$

Donde **U** es una matriz $n \times p$ que contiene los autovectores ortonormales de **B** asociados a los autovalores ordenados $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq \lambda_{p+1} = 0$, y **Λ** es la matriz diagonal de orden $p \times p$ que contiene los autovalores positivos $\lambda_1, \lambda_2, \dots, \lambda_p$.

Considerando esta descomposición espectral, tenemos lo siguiente: Consideremos la matriz $\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}$.

1.4.1 Solución por Coordenadas Principales

Es la matriz de coordenadas **X** tal que sus columnas $X_1, \dots, X_p$ (que interpretaremos como variables), son autovectores de **B** de los autovalores $\lambda_1, \lambda_2, \dots, \lambda_p$. Las coordenadas del elemento $i \in \Omega$ son $\mathbf{x}_\mathrm{i}' = (x_{i1}, \dots, x_{ip})$ donde $\mathbf{x}_\mathrm{i}'$ es la i-ésima fila de la matriz **X**, reciben el nombre de COORDENADAS PRINCIPALES y cumplen con que $\delta_{ij}^2 = \sum_{\alpha=1}^{p}(x_{i\alpha} - x_{j\alpha})^2 = (\mathbf{x}_\mathrm{i} - \mathbf{x}_\mathrm{j})'(\mathbf{x}_\mathrm{i} - \mathbf{x}_\mathrm{j})$, es decir, sus distancias son euclídeas.

Siendo **B** de orden $n \times n$, **U** es de orden $n \times p$, **Λ** es de orden $p \times p$, así **X** es de orden $n \times p$. En la práctica interesa representar el colectivo de individuos Ω, que lo visualizamos más fácilmente en el plano, por lo tanto, se toman las 2 primeras Coordenadas Principales donde las distancias son muy aproximadas y tenemos un mapa de los individuos. Las primeras $q < p$ Coordenadas Principales cuando $q = 2$ son análogas a una matriz de datos, pues no se obtuvieron los datos midiendo variables, si no a partir de distancias.

La Solución por Coordenadas Principales tiene importantes propiedades enunciadas a continuación.

Propiedades de X:

1. Las variables X_k con $k = 1, \dots, p$ tienen media 0, es decir, $\bar{x}_1 = \cdots = \bar{x}_p = 0$.
2. Las varianzas son proporcionales a los autovalores, es decir, $S_k^2 = \frac{1}{n}\lambda_k$ con $k = 1, \dots, p$. El coeficiente de proporcionalidad es el inverso de la numerosidad del colectivo.
3. Las variables están incorreladas, es decir, $Cor(X_i, X_j) = 0$ con $i \neq j = 1, \dots, p$. Así su matriz de correlaciones es la Identidad.
4. Las variables X_k son las Componentes Principales de cualquier matriz de datos **Z** tal que las distancias euclídeas entre sus filas concuerden con **Δ**. Esta propiedad es muy importante porque relaciona el Escalado Multidimensional con el Análisis de Componentes Principales.
5. La Variabilidad Geométrica de **Δ** es $V_\delta(\mathbf{X}) = \frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij}^2 = \frac{1}{n}\sum_{k=1}^{p}\lambda_k$.
6. La Variabilidad Geométrica en dimensión q es máxima cuando tomamos las q primeras Coordenadas Principales, es decir, $V_\delta(\mathbf{X})_q = \frac{1}{2n^2}\sum_{i,j=1}^{n}\delta_{ij_{(q)}}^2 = \frac{1}{n}\sum_{k=1}^{q}\lambda_k$. Por ejemplo, va a interesar la Variabilidad Geométrica en dimensión 2, y es máxima cuando tenemos las 2 primeras Coordenadas Principales, y así sucesivamente.

$\frac{\frac{1}{n}\sum_{k=1}^{q}\lambda_k}{\frac{1}{n}\sum_{k=1}^{p}\lambda_k}$ es la fracción de Variabilidad Geométrica. Y el porcentaje de variabilidad explicado por los q primeros ejes principales es la proporción de variabilidad geométrica $\frac{\sum_{k=1}^{q}\lambda_k}{\sum_{k=1}^{p}\lambda_k}$

1.4.2 Similaridad

En el campo de la Biología suele ser necesario estudiar similaridades, para lo cual se utiliza Escalado Multidimensional No Métrico. Se debe transformar a la matriz de similaridades en una matriz de distancias euclídeas.

Sea $\Omega = \{1, \dots, n\}$ un colectivo de numerosidad n. Una SIMILARIDAD sobre Ω es una función $S: \Omega \times \Omega \to \mathbb{R}$ tal que: $s(i,i) \geq s(i,j) = s(j,i) \geq 0$.

Suponiendo que tenemos p variables binarias $X_1, \dots, X_p$ donde cada X_i puede tomar valores 0 o 1 (dicotomías). Para cada par de individuos (i, j) consideramos la siguiente tabla:

		j	
		1	0
i	1	a	b
	0	c	d

Donde a, b, c, d son las frecuencias de (1,1), (1,0), (0,1) y (0,0) respectivamente, con $p = a + b + c + d$. Con esto se definen varios coeficientes de similaridad, por ejemplo:

Coeficiente de similitud de Jaccard

$$s_{ij} = \frac{a}{a + b + c}$$

Coeficiente de similitud de Dice

$$s_{ij} = \frac{2a}{2a + b + c}$$

Coeficiente de similitud de Sokal-Michener

$$s_{ij} = \frac{a + d}{p}$$

Estos verifican que $s_{ii} \geq s_{ij} = s_{ji} \geq 0$

La matriz de similaridades es $\mathrm{S} = \begin{pmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{pmatrix}$

Se puede transformar una similaridad en distancia aplicando la siguiente fórmula: $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$ (Paso de S a $\boldsymbol{\Delta}$).

Ahora se pueden considerar las matrices $\mathbf{A} = -\frac{1}{2}\boldsymbol{\Delta}^{(2)}$ y $\mathbf{B} = \mathbf{HAH}$. Se enuncian algunas propiedades:

1. Si S es una matriz (semi)definida positiva, la distancia d_{ij} es euclídea.
2. El rango de la matriz **H**S**H** es: $Rng(\mathbf{HSH}) = Rng(\mathrm{S}) - 1$.
3. Las Coordenadas Principales se obtienen diagonalizando $\mathbf{B} = \mathbf{HSH}$.

1.4.3 Escalado Multidimensional No Métrico

B no necesariamente suele ser semidefinida positiva, es decir que tiene autovalores negativos. En tal caso no se puede aplicar el Escalado Multidimensional Métrico, y hay que transformar las distancias. Suponiendo que $\boldsymbol{\Delta} = (\delta_{ij})$ sea una matriz de distancias NO euclídea, entonces **B** (de acuerdo con el teorema anterior) no es semidefinida positiva y, por tanto, tiene autovalores negativos: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0 \geq \lambda_{p+1} \geq \lambda_{p+2} \geq \cdots \geq \lambda_p'$. Como esto ocurre porque $\boldsymbol{\Delta}$ no es euclídea, el objetivo del Escalado Multidimensional No Métrico es hacer una transformación de esas δ_{ij} de manera que se vuelvan euclídeas, pero conservando las relaciones de proximidad entre los individuos del conjunto de n unidades estadísticas Ω. Para conservar las relaciones de proximidad se introduce el concepto de PRE-ORDENACIÓN.

Pre-ordenación: La pre-ordenación asociada a la matriz de distancias $\boldsymbol{\Delta}$ es la ordenación de las $m = \frac{n(n-1)}{2}$ (número de distancias debajo de la diagonal principal) distancias

$$\delta_{i_1 j_1} \leq \delta_{i_2 j_2} \leq \cdots \leq \delta_{i_m j_m}$$

La pre-ordenación es una propiedad asociada a Ω, es decir, podemos escribir: $(i_1, j_1) \preccurlyeq (i_2, j_2) \preccurlyeq \cdots \preccurlyeq (i_m, j_m)$[3]

$(i_k, j_k) \in \Omega \times \Omega$ (producto cartesiano) donde $(i, j) \preccurlyeq (i', j')$ si $\delta_{ij} \leq \delta_{i'j'}$

El objetivo es, entonces, representar a Ω en un espacio conservando lo más que se pueda la pre-ordenación.

Si transformamos δ_{ij} en $\hat{\delta}_{ij} = \varphi(\delta_{ij})$ donde φ es una función positiva y creciente, es evidente que se conserva la pre-ordenación original, por tanto, individuos próximos (o alejados) según δ_{ij} también serán próximos (o alejados) según $\hat{\delta}_{ij}$. Si además las distancias $\hat{\delta}_{ij}$ son euclídeas, ya se tiene la posibilidad de representar a Ω haciendo un Análisis de Coordenadas Principales, pero sobre la matriz de distancias $\widehat{\boldsymbol{\Delta}} = (\hat{\delta}_{ij})$ conservando aproximadamente la pre-ordenación. φ es una función generalmente no lineal y se la obtiene por regresión monótona. Una φ que se define a continuación es la transformación q-aditiva.

La TRANSFORMACIÓN Q-ADITIVA de δ_{ij} se define como:

$$\hat{\delta}_{ij}^{\,2} = \begin{cases} \delta_{ij}^{\,2} - 2a & si\ i \neq j \\ 0 & si\ i = j \end{cases}$$

Donde $a < 0$ es una constante. Esta transformación conserva la pre-ordenación.

Aún no se ha definido qué condiciones debe cumplir a para el Escalado Multidimensional. Se define a continuación bajo qué condiciones la transformación proporciona distancias euclídeas, mediante el siguiente teorema.

Teorema

Sea **Δ** una matriz de distancias NO euclídeas y sea $\lambda'_p < 0$ el menor autovalor de **B**, entonces la Transformación q-aditiva proporciona una distancia euclídea para todo a tal que $a \leq \lambda'_p$.

Al ser $a = \lambda'_p$, es decir, el último autovalor negativo, se distorsiona menos. De esta forma:

$$\widehat{\mathbf{A}} = \mathbf{A} - \lambda'_p(\mathbf{I} - \mathbf{J})$$

$$\widehat{\mathbf{B}} = \mathbf{B} - \lambda'_p\mathbf{H}$$

Y se puede demostrar que $\widehat{\mathbf{B}}$ tiene los mismos autovectores de **B**, y que sus autovalores son los autovalores de **B** restados a, es decir, $\hat{\lambda}_j = \lambda_j - a$.

[3] $\preccurlyeq$ se lee como "no sigue a" o "precede o es igual a" y es usado como relación entre parejas de individuos elementos de Ω.

1.5 Transformación Procrustes

Es una transformación que se aplica a una matriz o configuración $\mathbf{X}_2$ con el fin de que las discrepancias con respecto a una matriz o configuración de referencia $\mathbf{X}_1$ sean mínimas, para lo cual se emplea un factor de escala ρ y una matriz de transformación ortogonal $\mathbf{H}$, siendo la transformación: $\mathbf{Z} = \rho\mathbf{X}_2\mathbf{H}$. La idea es que ρ y $\mathbf{H}$ minimicen la suma de cuadrados de inter-distancias entre los puntos de ambas configuraciones que viene dada como: $tr[(\mathbf{X}_1 - \rho\mathbf{X}_2\mathbf{H})'(\mathbf{X}_1 - \rho\mathbf{X}_2\mathbf{H})]$. Si las matrices no tienen la misma dimensión, la de dimensión más baja deberá completarse con ceros hasta alcanzar la dimensión de la otra. Sin pérdida de generalidad todas las configuraciones se supone que están centradas para evitar el problema de la traslación, para lo cual se multiplica a cada matriz por la matriz de centrado. (Zuliani P., 2012, pp. 68-70)

1.5.1 Rotación Procrustes

Se impone la restricción de que $\mathbf{H}$ sea ortogonal para que la transformación sea una rotación o una reflexión. La expresión que minimiza la traza mencionada es: $\mathbf{H} = \mathbf{U}_2\mathbf{U}_1'$, siendo $\mathbf{U}_1$ y $\mathbf{U}_2$ las matrices de vectores propios de $\mathbf{X}_1'\mathbf{X}_2\mathbf{X}_2'\mathbf{X}_1$ y $\mathbf{X}_2'\mathbf{X}_1\mathbf{X}_1'\mathbf{X}_2$, respectivamente. Sólo se consideran los autovalores no negativos para garantizar que la $tr(\mathbf{X}_2\mathbf{H}\mathbf{X}_1')$ sea lo más grande posible. (Zuliani P., 2012, pp. 68-70)

1.5.2 Escalamiento

Para que las escalas de $\mathbf{X}_1$ y $\mathbf{X}_2\mathbf{H}$ coincidan y sean comparables es necesario dilatar o comprimir mediante el escalar ρ, el cual viene dado como: $\rho = \frac{tr(\mathbf{H}'\mathbf{X}_2'\rho\mathbf{X}_2\mathbf{H}\mathbf{X}_1)}{tr(\mathbf{X}_2'\rho\mathbf{X}_2\mathbf{H}\mathbf{X}_1)}$.

La rotación y el escalamiento pueden ser aplicados de forma independiente. (ZULIANI 2012)

1.6 Análisis Procrustes Generalizado

Sean m matrices $\mathbf{X}_k$ cada una con p_i variables y los mismos n individuos. La técnica consiste en minimizar la Suma de Cuadrados entre individuos (puntos análogos) después de haber sido rotados, trasladados y escalados, con la finalidad de conseguir un nuevo espacio que represente las diferencias entre individuos, manteniendo las distancias entre ellos (CALDERÓN CISNEROS 2016). Se reproduce a continuación la introducción de la explicación de esta técnica, proporcionada por Gower (GOWER, Generalized Procrustes Analysis 1975):

Supónganse que $P_i^{(k)}$ con $k = 1, \dots, m$ y $i = 1, \dots, n$ representan las localizaciones de mn puntos en el espacio p dimensional. De manera colectiva éstas se pueden considerar como m configuraciones, cada una con n puntos en p dimensiones. Se trata de trasladar, rotar/reflectar y escalar las m configuraciones para minimizar el criterio de bondad de ajuste:

$$\sum_{k=1}^{m} \sum_{i=1}^{n} \Delta^2 (P_i^{(k)}, G_i)$$

Donde G_i es el centroide de los m puntos $P_i^{(k)}$. Las posiciones rotadas de cada configuración k se pueden considerar como un análisis individual con la configuración del centroide representando un consenso. La técnica para calcularlo se puede resumir en forma de análisis de varianza. El caso especial donde $m = 2$ corresponde al Análisis Procrustes clásico, pero a diferencias de éste la elección del criterio que ajusta cada configuración al centroide común evita las dificultades que surgen cuando una configuración se ajusta a la otra que se queda fija (Ajustar $\boldsymbol{X}_2$ a $\boldsymbol{X}_1$ no da el escalado inverso de ajustar $\boldsymbol{X}_1$ a $\boldsymbol{X}_2$).

Supónganse que $\boldsymbol{X}_k$ con $k = 1, \dots, m$ es una matriz con n filas y p_i columnas, cuya i-ésima fila da las coordenadas de un punto $P_i^{(k)}$, referido en p_k ejes ortogonales.

A partir de ahora se asume que los m puntos $P_i^{(k)}$ $(k = 1, \dots, m)$ se refieren a la misma i-ésima entidad. Por ejemplo, cada $\boldsymbol{X}_k$ puede haber sido obtenida desde m diferentes conjuntos de variables observadas para las mismas muestras en cada caso (cada muestra enumerada por i), o cada $\boldsymbol{X}_k$ se podría haber obtenido como diferentes escalas de los mismos datos experimentales (cada experimento enumerado por i), o cada $\boldsymbol{X}_k$ se podría haber levantado desde un tipo de escala de un mismo estímulo según lo percibido por diferentes individuos (cada estímulo enumerado por i). También se asume que $p_k = p$, una constante, pues esto simplifica la explicación sin pérdida de generalidad porque basta con escoger $p = Max(p_k)$ y añadir columnas de ceros a cada $\boldsymbol{X}_k$ que inicialmente tenga menos columnas que p.

Son comunes los problemas donde se desea estudiar las relaciones entre los m conjuntos y con frecuencia se desea algún tipo de análisis combinado.

La idea de Procrustes es generalizar que las m configuraciones sean simultáneamente trasladados, rotados, reflectados y escalados de tal manera que un criterio de bondad de ajuste sea optimizado. El criterio adoptado es minimizar la suma de cuadrados entre cada grupo de m puntos análogos $P_i^{(k)} (k = 1, \dots, m)$ y su centroide G_i sumado a todos los n grupos. Los centroides se muestran en la **Figura 3-1** *para $m = 3$ configuraciones, $p_1 = p_2 = p_3 = 2$ dimensiones, referidos a las mismas $n = 4$ entidades, y el centroide de todo el sistema está en 0:*

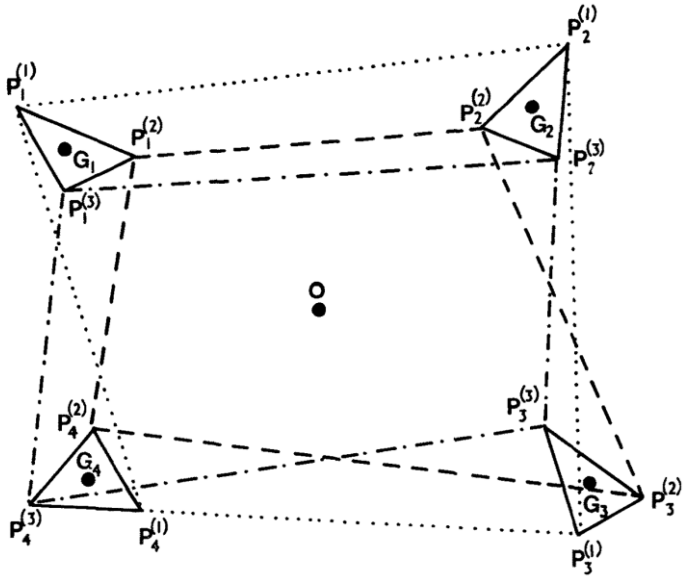

Figura 1-1: APG. Ubicación espacial de los puntos y sus centroides (3 grupos, 2 variables, 4 individuos).

Fuente: *(GOWER, Generalized Procrustes Analysis 1975)*

Las mn longitudes $\Delta^2(P_i^{(k)}, G_i)$ *son denominadas residuos. La suma de cuadrados de los residuos* S_r *es, por lo tanto:*

$$S_r = \sum_{i=1}^{n}\sum_{k=1}^{m} \Delta^2(P_i^{(k)}, G_i)$$

Sin embargo, por la identidad:

$$\sum_{u<v}^{m} \Delta^2(P_i^{(u)}, P_i^{(v)}) \equiv m \sum_{u=1}^{m} \Delta^2(P_i^{(u)}, G_i)$$

Es más simple trabajar en términos de

$$S = \sum_{i=1}^{n}\sum_{u<v}^{m} \Delta^2(P_i^{(u)}, P_i^{(v)})$$

En lugar de minimizar una suma de cuadrados de residuos, podría ser considerado otro criterio que podría tener propiedades deseables. Como es usual el criterio de mínimos cuadrados conduce a álgebra manejable y cálculos sencillos.

El problema de minimización es necesario trabajarlo de forma algebraica. Se tienen m *matrices de orden* $n \times p$ *denotadas con* $\boldsymbol{X}_k$ *con* $k = 1, \dots, m$ *donde la* i*-ésima fila de* $\boldsymbol{X}_k$ *se interpreta como las coordenadas dadas de un punto* $P_i^{(k)}$ *en el espacio Euclidiano. Rotar las configuraciones de n puntos dadas por* $\boldsymbol{X}_k$ *es equivalente a post-multiplicar* $\boldsymbol{X}_k$ *por una matriz ortogonal* $\boldsymbol{H}_k$*, y un escalado uniforme es expresado por una constante multiplicativa* ρ_k*. La traslación a un nuevo origen se logra añadiendo el mismo vector fila* $1 \times p$ $\boldsymbol{t}_k$ *a cada fila de* $\boldsymbol{X}_k$*. Escribiendo* $\boldsymbol{T}_k$ *para la matriz* $n \times p$*, cuyas filas son* $\boldsymbol{t}_k$*, luego escalando, rotando y trasladando se expresan algebraicamente por la transformación*

$$\boldsymbol{X}_k \rightarrow \rho_k \boldsymbol{X}_k \boldsymbol{H}_k + \boldsymbol{T}_k$$

Se requiere determinar un ρ_k, $\boldsymbol{H}_k$, $\boldsymbol{T}_k$ $(k = 1, \dots, m)$ *tal que la Suma de Cuadrados de los Residuos* S_r *sea mínima. Así el problema algebraico es minimizar:*

$S \equiv tr \sum_{u<v}^{m} [(\rho_u \boldsymbol{X}_u \boldsymbol{H}_u + \boldsymbol{T}_u) - (\rho_v \boldsymbol{X}_v \boldsymbol{H}_v + \boldsymbol{T}_v)] \cdot [(\rho_u \boldsymbol{X}_u \boldsymbol{H}_u + \boldsymbol{T}_u) - (\rho_v \boldsymbol{X}_v \boldsymbol{H}_v + \boldsymbol{T}_v)]'$ (GOWER, Generalized Procrustes Analysis 1975)

Se reproduce y resume a continuación una esquematización realizada por Zuliani (ZULIANI 2012) de cómo deben ser los valores para la transformación de cada matriz $\mathbf{X}_k$, y a qué restricciones se sujeta cada uno.

1.6.1 Traslación

Se prueba que la Suma de Cuadrados Residual S_r se minimiza cuando los vectores $\mathbf{t}_k$ que componen las filas de $\mathbf{T}_k$ son iguales para las m configuraciones: $\mathbf{t}_1 = \mathbf{t}_2 = \cdots = \mathbf{t}_m$, lo que implica que las m configuraciones deberían ser trasladadas para tener el mismo centroide, el cual, para simplificar cálculos y sin pérdida de generalidad, conviene que sea el origen. Esto se logra multiplicando cada matriz por la matriz de centrado. (ZULIANI 2012)

Luego de esto se sugiere una normalización, estandarización o escalamiento inicial si existen desproporciones inusuales en las magnitudes de los atributos. (ZULIANI 2012)

1.6.2 Rotación/Reflexión

La matriz de rotación $\mathbf{H}_k$ tiene la restricción de ser ortogonal. Siendo $\mathbf{Y} = \frac{1}{m} \sum_{k=1}^{m} \rho_k \mathbf{X}_k \mathbf{H}_k$ las coordenadas del centroide del grupo o configuración consenso después de la rotación y escalamiento, se escribe $\rho_k \mathbf{X}_k' \mathbf{Y}$ en sus valores singulares de la forma $\mathbf{U}_k \mathbf{\Gamma}_k \mathbf{V}_k'$ (con $\mathbf{U}_k$y $\mathbf{V}_k$ ortogonales y $\mathbf{\Gamma}_k$ diagonal) y notando que ésta es una matriz simétrica, resulta: $\mathbf{H}_k = \mathbf{U}_k \mathbf{V}_k'$. Es decir que la rotación que minimiza la Suma de Cuadrados Residual es una rotación procrusteana con matriz objetivo la configuración consenso $\mathbf{Y}$. Si en la matriz de rotación se presentan valores negativos es indicio de que se está realizando una reflexión. (ZULIANI 2012)

Hasta esta etapa además de minimizar la distancia entre puntos homólogos de las configuraciones originales, se están preservando las distancias entre individuos. (ZULIANI 2012)

1.6.3 Escalamiento

Es el último paso. Se puede decidir no escalar, pero si se ha decidido hacerlo se puede hacer para expandir/estirar (factor de escala > 1) o constreñir/contraer (factor de escala < 1) las distancias de los puntos de la configuración k al origen para acercarse a la configuración objetivo. Si se escala, las variaciones entre configuraciones asociadas a diferentes criterios de escalas empleadas en las configuraciones desaparecen. Para minimizar la Suma de Cuadrados Residual y hallar los factores de escala óptimos (sin la solución trivial de que todos los ρ_k sean igual a 0) la restricción es: $\sum_{k=1}^{m} \rho_k^2 tr(\mathbf{X}_k \mathbf{X}_k') = \sum_{k=1}^{m} tr(\mathbf{X}_k \mathbf{X}_k')$. De esta forma la Suma de Cuadrados Residual permanece invariante se incluyan o no los parámetros ρ_k, ya que no siempre es necesario incluirlos. Se encuentra que: $\rho_k = \frac{tr(\mathbf{X}_k \mathbf{H}_k \mathbf{Y}') \sum_{k=1}^{m} tr(\mathbf{X}_k \mathbf{X}_k')}{m\, tr(\mathbf{X}_k \mathbf{X}_k')\, tr(\mathbf{Y}\mathbf{Y}')}$. (ZULIANI 2012)

Así se establecen los parámetros que minimizan la Suma de Cuadrados Residual, pero como **Y** contiene a $\mathbf{H}_k$ y ρ_k, no se puede calcular directamente, por lo que se calcula por medio de un proceso iterativo. (ZULIANI 2012)

1.6.4 Proceso iterativo

Una vez estandarizadas y trasladadas al origen las m configuraciones, en la primera iteración se las rota y escala. Después de la primera iteración se calcula la configuración consenso **Y** como la media de las configuraciones transformadas. Con esto se inicia la segunda iteración de rotación y escalamiento fijando a **Y**, y así sucesivamente hasta que el cambio en la Suma de Cuadrados Residual entre iteraciones sucesivas se menor que un valor de tolerancia fijado (por ejemplo 0.0001).

Se reproduce el resumen de los cálculos, extraído de Gower (GOWER, Generalized Procrustes Analysis 1975):

1. *Evaluar la Suma de Cuadrados Entre grupos (es decir el término de traslación en el Análisis de Varianza).*
2. *Centrar cada $\boldsymbol{X}_k$ y escalar cada $\boldsymbol{X}_k$ por λ de modo que $\sum_{k=1}^{m} \lambda\, tr(\boldsymbol{X}_i \boldsymbol{X}_i') = m$.*
3. *Establecer $\boldsymbol{Y} = \boldsymbol{X}_1$ (ajuste inicial de la matriz de medias). Para $k = 2, 3, \dots, m$ rotar $\boldsymbol{X}_k$ para ajustar a $\boldsymbol{Y}$, y reevaluar $\boldsymbol{Y}$ como la media de $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_k$. Evaluar la Suma de Cuadrados de Residuos inicial $S_r = m(1 - tr(\boldsymbol{Y}\boldsymbol{Y}'))$ y establecer a $\rho_k = 1$ para $k = 1, \dots, m$.*
4. *Para $k = 1, \dots, m$ rotar la matriz actual $\rho_k \boldsymbol{X}_k$ para ajustar a $\boldsymbol{Y}$ mediante $\boldsymbol{X}_k^* = \rho_k \boldsymbol{X}_k \boldsymbol{H}_k$. Calcular $\boldsymbol{Y}^*$ y $S_r^* = S_r - m\, tr(\boldsymbol{Y}^* \boldsymbol{Y}^{*\prime} - \boldsymbol{Y}\boldsymbol{Y}')$. Establecer $S_r^{**} = S_r^*$.*
5. *Si no se requiere escalado, ir al paso 7.*

6. *Para* $k = 1, \dots, m$ *evaluar* $\frac{\rho_k^*}{\rho_k}$ *desde* $\frac{\rho_k^{*2}}{\rho_k^2} = \frac{tr(\rho_k \boldsymbol{X}_k \boldsymbol{Y}')}{tr(\rho_k^2 \boldsymbol{X}_k \boldsymbol{X}_k')tr(\boldsymbol{Y}\boldsymbol{Y}')}$, *escalar* $\boldsymbol{X}_k^{**} = \frac{\rho_k^*}{\rho_k}\boldsymbol{X}_k^*$ *y establecer* $\rho_k = \rho_k^*$. *Calcular la nueva media* $\boldsymbol{Y}^{**}$ *y la Suma de Cuadrados Residual* $S_r^{**} = S_r^* - m\ tr\left(\boldsymbol{Y}^{**}\boldsymbol{Y}^{**\prime} - \boldsymbol{Y}^*\boldsymbol{Y}^{*\prime}\right)$.
7. *Si* $S_r - S_r^{**} > tolerancia$, *establecer* $S_r - S_r^{**}$ *e ir al paso 4, caso contrario ir al siguiente paso.*
8. *La iteración se ha completado. Calcular e imprimir el Análisis de Varianza.* (GOWER, Generalized Procrustes Analysis 1975)

Los valores de la configuración consenso encontrada en la última iteración pueden ser utilizados para construir dendrogramas o para realizar ACP. (ZULIANI 2012)

Tabla 1-1: Análisis de Varianza Procrustes (PANOVA)

Entre grupos		Traslación	
Dentro de grupos Por individuo	Consenso	Residual	Total (Dentro de grupos)
1	$m\Delta^2(O, G_1)$	$\sum_{k=1}^{m} \Delta^2(P_1^{(k)}, G_1)$	$\sum_{k=1}^{m} \Delta^2(O, P_1^{(k)})$
2	$m\Delta^2(O, G_2)$	$\sum_{k=1}^{m} \Delta^2(P_2^{(k)}, G_2)$	$\sum_{k=1}^{m} \Delta^2(O, P_2^{(k)})$
.	.	.	.
.	.	.	.
.	.	.	.
n	$m\Delta^2(O, G_n)$	$\sum_{k=1}^{m} \Delta^2(P_n^{(k)}, G_n)$	$\sum_{k=1}^{m} \Delta^2(O, P_n^{(k)})$
Individuo	$m\ tr(\mathbf{YY}')$	S_r	$\sum_{k=1}^{m} tr(\mathbf{X}_k\mathbf{X}_k')$
Por grupo	Consenso	Residual	Total (Dentro de grupos)
1		$\sum_{i=1}^{n} \Delta^2(P_i^{(1)}, G_i)$	$\sum_{i=1}^{n} \Delta^2\left(O, P_i^{(1)}\right) = \rho_1^2\ tr(\mathbf{X}_1\mathbf{X}_1')$
2		$\sum_{i=1}^{n} \Delta^2(P_i^{(2)}, G_i)$	$\sum_{i=1}^{n} \Delta^2\left(O, P_i^{(2)}\right) = \rho_2^2\ tr(\mathbf{X}_2\mathbf{X}_2')$
.		.	.
.		.	.
.		.	.
m		$\sum_{i=1}^{n} \Delta^2(P_i^{(m)}, G_i)$	$\sum_{i=1}^{n} \Delta^2\left(O, P_i^{(m)}\right) = \rho_m^2\ tr(\mathbf{X}_m\mathbf{X}_m')$
Grupos		S_r	$\sum_{k=1}^{m} tr(\mathbf{X}_k\mathbf{X}_k')$

Fuente: (GOWER, Generalized Procrustes Analysis 1975)

1.6.5 Análisis de Varianza Procrustes (PANOVA)

Lo anterior puede ser expresado en la forma de un Análisis de Varianza que es útil para identificar la importancia relativa de los elementos que van a contribuir con la Suma de Cuadrados Total, que se puede descomponer en *Entre grupos* y *Dentro de grupos*, entendiéndose como grupos a

las m configuraciones. La componente de *Entre grupos* representa la contribución de los términos de traslación. De forma univariada la bien conocida identidad de Análisis de Varianza es: $\sum_{k=1}^{m} y_k^2 \equiv m\bar{y}^2 + \sum_{k=1}^{m}(y_k - \bar{y})^2$. De manera análoga la forma multivariada es: $tr\left(\sum_{k=1}^{m} \rho_k^2 \mathbf{X}_k \mathbf{X}_k'\right) \equiv m\, tr(\mathbf{YY}') + S_r$, en donde el lado izquierdo es la Suma de Cuadrados *Dentro de grupos* después del escalado y el rotado (y gracias a la restricción del factor de escala, es el mismo valor que la Suma de Cuadrados *Dentro de grupos* antes de la transformación). El k-ésimo término es la contribución de la k-ésima configuración a la Suma de Cuadrados Total *Dentro de grupos*, y muestra la reducción, o incremento, debido al escalado. En el lado derecho, de la identidad $m\, tr(\mathbf{YY}')$ es el término que representa la contribución de la configuración promedio o consenso. Geométricamente esto es $m \sum_{i=1}^{n} \Delta^2(O, G_i)$. El residual S_r se puede partir de dos formas: la una sumando los cuadrados de los residuos para cada grupo, dando los términos $\sum_{i=1}^{n} \Delta^2(P_i^{(k)}, G_i)$, y la otra sumando los cuadrados para cada fila dando los términos $\sum_{k=1}^{m} \Delta^2(P_i^{(k)}, G_i)$ (GOWER, Generalized Procrustes Analysis 1975) (ZULIANI 2012). Estas consideraciones dan la forma tabular del Análisis de Varianza expresado geométricamente en la **Tabla 1-1**.

1.7 Marcador molecular

Koebner y Centre (KOEBNER y CENTRE 2003) definen los marcadores moleculares de la siguiente manera:

> *Los marcadores moleculares son ensayos de diagnóstico de genotipo, y se basan específicamente en la variación a nivel del ADN. Cada marcador define un segmento de ADN particular dentro de un genoma.*
>
> *Los marcadores moleculares representan un subconjunto de diagnósticos de genotipo (o "marcadores genéticos"), específicamente los que se basan en un análisis directo de ADN, en lugar de, por ejemplo, en un rasgo fenotípico heredado (p. Ej., Color de flor en guisante, Pisum sativum, o semi-enanismo en cereales, conocidos como "marcadores morfológicos") o un ensayo basado en proteínas de un producto genético (por ejemplo, amilasa o proteínas de almacenamiento de semillas - "marcadores bioquímicos"). Definido estrictamente, un marcador molecular identifica un segmento de ADN genómico, dentro del cual la variación alélica en secuencia ha permitido que su posición se mapee genéticamente (…). Su significado menos definido, pero mucho más ampliamente utilizado, se extiende a un ensayo basado en ADN de un locus[4], esté o no genéticamente ligado a cualquier gen particular de interés.* (KOEBNER y CENTRE 2003)

Existen varios tipos de marcadores moleculares que permiten hacer los estudios a nivel molecular. Algunos que se han utilizado habitualmente son: los RFLP (Restriction Fragment Length Polymorphism o Polimorfismos de longitud de fragmentos de restricción), AFLP (Amplified Fragment Length Polymorphisms o Polimorfismo de Longitud de Fragmento Amplificado) y

[4] locus / loci: (singular / plural) posición fija en el cromosoma (ubicación física identificable).

RAPD (Random Amplification of Polymorphic DNA o Amplificación Aleatoria de ADN Polimórfico). Uno de los métodos más usados para caracterización genética, estudios de diversidad y programas de selección asistida son los marcadores Microsatélites SSR (Simple Sequence Repeats o Repeticiones de Secuencia Simple) (LOZADA VARGAS 2014). A partir de un foro internacional, representado por académicos, gobiernos y científicos de la industria del cacao, los análisis de ADN por Microsatélites fueron descritos como la herramienta molecular más apropiada para la identificación molecular del cacao (LOOR SOLORZANO 2002).

1.7.1 Microsatélites SSR

Vázquez O., et al. (VAZQUEZ OVANDO, y otros 2012) enuncian que:

> *Las secuencias simples repetidas (SSR) por su nombre en inglés, Simple Sequence Repeat o microsatélites son segmentos cortos de ADN (de 1 a 10 pb), que se repiten en serie y de forma aleatoria a través de todo el genoma de los seres vivos. (...) Poseen varias ventajas frente a otros marcadores (minisatélites, RFLP, RAPD, AFLP, etc.) ya que son muy polimórficos. (...) Estas regiones son a menudo altamente variables y consecuentemente útiles para medir el polimorfismo entre especies o variedades muy relacionadas, característica ampliamente deseable en los estudios poblaciones para evaluar la diversidad genética. (...) Debido a estas importantes ventajas, los microsatélites han sido el método de elección para conducir investigaciones en cacao en la última década.* (VAZQUEZ OVANDO, y otros 2012)

Detectan un alto nivel de polimorfismo (múltiples alelos de un gen), y son codominantes (AZOFEIFA DELGADO 2006).

1.8 Análisis de Varianza Molecular (AMOVA)

Ruiz E. (2014) señala que el AMOVA:

> *Permite estudiar la variación molecular dentro de una especie, y se basa en un modelo jerárquico anidado. Se puede atribuir un porcentaje de variación Entre grupos, Dentro de grupos e Individual dentro de grupos. El modelo es:*
>
> $$Y_{ki(j)} = Y + A_k + B_{k(i)} + W_{ki(j)}$$
>
> *Donde A_k es el efecto de la k-ésima población con varianza σ_a^2, $B_{k(i)}$ es el efecto del i-ésimo individuo dentro de la k-ésima población, con varianza σ_b^2, $W_{ki(j)}$ es el efecto del j-ésimo locus del i-ésimo individuo de la k-ésima población, con varianza σ_w^2.* (RUIZ ERAZO 2014)

1.9 Coeficiente de similaridad apropiado para marcadores moleculares SSR y organismos diploides

Kosman y Leonard (KOSMAN y LEONARD 2005) aclaran que para elegir un coeficiente de similitud apropiado para medir adecuadamente la disimilaridad genética a partir de marcadores moleculares es importante considerar dos aspectos: 1. La ploidía o número de cromosomas de la especie en estudio, y 2. La dominancia del marcador molecular empleado. Es importante considerar estos aspectos debido a que los coeficientes de similitud que son usualmente empleados como Dice, Jaccard y Sokal-Michener, son más apropiados para haploides. Afirman que tampoco es tan obvio que el coeficiente de similaridad de Nei & Li (1979) tan ampliamente recomendado sea siempre adecuado para comparar perfiles de huella digital (KOSMAN y LEONARD 2005). Por este motivo para diploides con marcadores codominantes Kosman y Leonard (KOSMAN y LEONARD 2005) desarrollaron una medida de disimilaridad intralocus (que considera estados homocigotos y heterocigotos) que se expande para medir la disimilaridad entre los estados multilocus de dos individuos, promediándola a través de todos los loci codominantes, y que cumple con ser una métrica.

La matriz generada por los marcadores moleculares contiene n individuos diploides que conforman las filas ($i = 1, \dots, n$), cada uno de los cuales ha sido sujeto a análisis genético con marcadores moleculares en p loci (es decir con p marcadores moleculares). Cada locus multialélico se denota con f_j ($j = 1, \dots, p$) y se denota con r_j al número de alelos que aparecieron en el locus j. Se denota con $S_j = \{s_{j1}, s_{j2}, \dots, s_{jr}\}$ al conjunto de estos alelos en el locus j. Cada individuo i se representa por su vector de estados $i = \{t_1, t_2, \dots, t_p\}$ donde t_j representa el estado del locus j, $t_j =< s_{ju}, s_{jv} >$, sea éste homocigoto ($s_{ju} = s_{jv}$) o heterocigoto ($s_{ju} \neq s_{jv}$) con $u, v = 1, \dots, r_j$. (KOSMAN y LEONARD 2005)

Para este enfoque, suponiendo 4 alelos A, B, C y D en el locus j, es decir $r_j = 4$ y $S_j = \{A, B, C, D\}$ se define que:

Tabla 2-1: Porcentaje de identidad entre 2 individuos de acuerdo a sus estados.

Estado del individuo i_1	Estado del individuo i_2	Identidad entre i_1 e i_2
$t_j =< A, A >$ Homocigoto	$t_j =< A, A >$ Homocigoto	100%
$t_j =< A, B >$ Heterocigoto	$t_j =< A, B >$ Heterocigoto	100%
$t_j =< A, A >$ Homocigoto	$t_j =< A, B >$ Heterocigoto	50%
$t_j =< A, B >$ Heterocigoto	$t_j =< A, C >$ Heterocigoto	50%
$t_j =< A, A >$ Homocigoto	$t_j =< B, B >$ Homocigoto	0%
$t_j =< A, A >$ Homocigoto	$t_j =< B, C >$ Heterocigoto	0%
$t_j =< A, B >$ Heterocigoto	$t_j =< C, D >$ Heterocigoto	0%

Realizado por: Gabriela J. Obregón O. 2018

Fuente: (KOSMAN y LEONARD 2005)

La similaridad entre estados para 3 alelos A, B y C se resume en la matriz:

	A,A	B,B	C,C	$A,B = B,A$	$A,C = C,A$	$B,C = C,B$
A,A	1	0	0	1/2	1/2	0
B,B		1	0	1/2	0	1/2
C,C			1	0	1/2	1/2
$A,B = B,A$				1	1/2	1/2
$A,C = C,A$					1	1/2
$B,C = C,B$						1

Fuente: (KOSMAN y LEONARD 2005)

Es decir, si comparten ambos alelos dentro del mismo locus, la similitud $s_{i_{1j}i_{2j}} = 1$, caso contrario, si comparten sólo uno $s_{i_{1j}i_{2j}} = 1/2$, caso contrario (no comparten ninguno), la similitud es $s_{i_{1j}i_{2j}} = 0$.

Una vez calculada las similaridades entre los individuos dentro de cada uno de los p loci, la similaridad final $s_{i_1 i_2}$ entre el individuo i_1 e i_2 se calcula promediando las similaridades genéticas intralocus, es decir, $s_{i_1 i_2} = \frac{1}{p}\sum_{j=1}^{p} s_{i_{1j}i_{2j}}$.

CAPÍTULO II

2 MARCO METODOLÓGICO

Como se mencionó en la sección *Antecedentes* en *Introducción*, Bramardi et. al, y Bruno y Balzarini, recomiendan Análisis Procrustes Generalizado (APG) para analizar datos morfológicos y moleculares llegando a un consenso entre ellos, lo que permite tomarlos en cuenta simultáneamente y así obtener una matriz de datos a partir de la cual se puede lograr una mejor clasificación de los individuos. Esta técnica al ser descriptiva no requiere del cumplimiento de supuestos.

Gower y Dijksterhuis (GOWER y DIJKSTERHUIS, Procrustes Problems 2004) establecen que, para emplear APG, las matrices de partida pueden estar formadas por observaciones directas de conjuntos de variables, donde se pueden ver los casos (filas) como un conjunto de puntos en un espacio euclídeo multidimensional siendo cada variable una dimensión, y cuya nube de puntos se denomina configuración, y se trata de hacer coincidir configuraciones. Otra opción es que las configuraciones sean de escalado multidimensional derivado de matrices de distancia cuyos elementos derivan de las variables originales. Todo esto siempre y cuando los casos en todas las configuraciones sean los mismos para que tenga sentido compararlos. Con estas consideraciones distinguen 3 principales tipos de datos en los que se puede aplicar la técnica: 1. Coordenadas derivadas de algún tipo de escalado multidimensional (o coordenadas de puntos de referencia directamente medidos si se trata de análisis de forma), 2. Matrices de datos cuyas columnas son diferentes variables, 3. Matrices de cargas derivadas de análisis factorial. (GOWER y DIJKSTERHUIS, Procrustes Problems 2004). Dependiendo del caso el número de variables de cada configuración puede o no ser el mismo, y las variables pueden ser comunes en las diferentes configuraciones o pueden no serlo.

Además, recomiendan que no haya incompatibilidades en la escala de los datos entre variables, y en caso de haberlas, existe la necesidad de eliminar los efectos con algún tipo de estandarización (tal como se hace con Análisis de Componentes Principales). Este problema se presenta principalmente cuando las configuraciones derivan directamente de los datos sin procesar, pero cuando derivan de escalado multidimensional, este aspecto es más sencillo.

Cuando se trata de matrices de distancia, se pueden calcular también a partir de información categórica empleando el coeficiente de distancia/disimilaridad/similaridad más apropiado.

En este estudio se recolectaron datos de las variables de interés, con un tamaño de muestra recomendado para cada unidad de análisis en cada una de las configuraciones, siendo éstas 4 de tipo morfológico y 1 de tipo molecular. Luego del tratamiento inicial recomendado en el cual se partió de las matrices de distancia para aplicar escalado multidimensional, se llegó a un consenso entre los dos tipos de variables, a partir del cual se obtuvieron grupos según las distancias entre clones. Se presentan los detalles en los siguientes puntos.

2.1 Hipótesis y especificación de las variables

En este estudio no se planteó una hipótesis estadística, en su lugar se plantearon preguntas de investigación. Se presenta el panorama general introductorio mediante la matriz de consistencia de la **Tabla 1-2**:

Tabla 1-2: Matriz de consistencia

Preguntas de investigación	**Objetivos**	**Variables en estudio**	**Unidad de análisis, población y muestra**
Pregunta general: ¿Cómo se van a asemejar los clones entre sí en el consenso entre las caracterizaciones molecular y morfológica?	General: Conocer el grado de similitud entre los 26 clones de cacao de tipo Criollo y los 4 clones testigo, considerando simultáneamente marcadores morfológicos y marcadores moleculares.	Marcadores morfológicos: Mazorca: 13 variables cuantitativas. Semilla: 3 variables cuantitativas y 1 cualitativa Hoja: 7 variables cuantitativas y 1 cualitativa Flor: 10 variables cuantitativas y 3 cualitativas Total: 38 variables	Unidad de análisis: hojas, mazorcas, semillas, flores, clones. Población: 26 clones de cacao tipo Criollo, 4 clones testigo.
Preguntas específicas: ¿Cómo se van a asemejar los clones entre sí sólo en la caracterización morfológica? ¿Cómo se van a asemejar sólo en la caracterización molecular? ¿Se agruparán de forma similar a como se agruparon en la caracterización morfológica, o será una agrupación diferente?	Específicos: • Preparar matrices de datos, obtener matrices de distancias/similaridades por clon. • Obtener solución por coordenadas principales de cada matriz de distancia/similaridad, comparar similitud entre clones. • Llegar a un consenso entre tipos de marcadores con	Marcadores moleculares: 20 Marcadores Microsatélites SSR.	Muestra: 30 hojas, 20 flores, 20 mazorcas y 100 semillas por mazorca recolectados de entre varios árboles para cada uno de los clones.

	Análisis Procrustes Generalizado, evaluar similitud entre clones. • Determinar conglomerados con agrupamiento jerárquico de Ward a partir de configuraciones individuales y consenso, comparar grupos de clones que se forman en cada uno.		

Realizado por: Gabriela J. Obregón O. 2018

2.2 Tipo y diseño de investigación

De acuerdo con Hernández S. et al. (HERNANDEZ SAMPIERI, FERNANDEZ COLLADO y BAPTISTA LUCIO 2010) esta investigación es cuantitativa y tiene un alcance primeramente descriptivo, al recoger información y medir las variables en estudio e indicar sus tendencias para proporcionar un panorama de los individuos mediante Análisis Procrustes Generalizado, técnica descriptiva (no inferencial). Sin embargo, también tiene un componente inferencial: en el caso de las variables morfológicas se pretende sacar conclusiones para una población a partir de una muestra en lo que respecta a la media, si se considera como población a las unidades de análisis de cada planta original (mazorca, semilla, hoja y flor) y de los clones que de ella se obtengan y, como muestra, el número de unidades de análisis que se necesita en cada variable, tomadas de diferentes clones de dicha planta (al ser genéticamente idénticos la teoría establece que la variación entre clones de una planta, bajo prácticamente las mismas condiciones ambientales, por estar plantados uno junto a otro, es mínima), para así caracterizar la planta; esto se hace con cada una de las plantas en estudio: EEB-1, EEB-2, etc. En el caso de las variables moleculares no hace falta hacer esta inferencia, pues los clones de una planta son genéticamente idénticos, es decir, a nivel de ADN no hay variabilidad entre éstos.

En segundo lugar, el alcance es correlacional, pues se conocerá la relación o grado de asociación entre las variables de caracterización morfológica y de caracterización molecular o cómo se comportan las primeras con respecto a las segundas, además de la correlación lineal entre variables dentro del tipo de caracterización. El diseño de la investigación es no experimental. El estudio es transversal.

2.3 Población de estudio

La población de interés son 26 distintos clones de cacao con características del grupo genético Criollo que proceden de Esmeraldas, presentados en la **Tabla 2-2**, y 4 clones testigo de grupo genético conocido: EET-103, EET-111, EET-116 y CCN-51, que proceden de Guayas, Trinidad y Tobago, Perú y Naranjal, respectivamente, presentados en la **Tabla 3-2**.

Tabla 2-2: Clones con características de Criollo.

Número	Nombre del clon	Codificación	Número	Nombre del clon	Codificación
1	EEB-1	C1	14	EEB-14	C14
2	EEB-2	C2	15	EEB-15	C15
3	EEB-3	C3	16	EEB-16	C16
4	EEB-4	C4	17	EEB-17	C17
5	EEB-5	C5	18	EEB-18	C18
6	EEB-6	C6	19	EEB-19	C19
7	EEB-7	C7	20	EEB-20	C20
8	EEB-8	C8	21	EEB-21	C21
9	EEB-9	C9	22	EEB-22	C22
10	EEB-10	C10	23	EEB-23	C23
11	EEB-11	C11	24	EEB-24	C24
12	EEB-12	C12	25	EEB-25	C25
13	EEB-13	C13	26	EEB-26	C26

Realizado por: Gabriela J. Obregón O. 2018

Tabla 3-2: Clones testigo.

Nombre del clon	Grupo genético
EET-103	Nacional
EET-111 (ICS-95)	Trinitario
EET-116 (IMC-67)	Forastero Amazónico
CCN-51	Híbrido Triple: Cruce entre ISC95*IMC67*Canelos

Realizado por: Gabriela J. Obregón O. 2018

2.4 Unidad de análisis

Clones de plantas de cacao: hojas, mazorcas, semillas y flores.

2.5 Tamaño de muestra

Como ya se mencionó, se necesita una muestra para inferir sobre las características morfológicas de un mismo individuo o planta, y obtener una caracterización de cada una, por lo cual se estudió un número determinado de cada unidad de análisis.

Soria y Enríquez (CENTRO AGRONÓMICO Y TROPICAL DE INVESTIGACIÓN Y ENSEÑANZA (CATIE) 1981), además de determinar estadísticamente las características de cacao más útiles o con mayor valor descriptivo, determinaron un tamaño de muestra mínimo requerido para cada una, siguiendo lo indicado por Torrie y Steel (1960) en su libro *Principios y procedimientos de Estadística*. Bajo esta recomendación, que es con la que usualmente se trabaja en el Programa de Cacao para caracterizaciones morfológicas, los tamaños de muestra para caracterizar a cada clon diferente fueron:

Número de mazorcas por clon: 20. Total: 600 mazorcas

Número de semillas por mazorca: 5, y por clon: 100. Total: 3000 semillas

Número de hojas por clon: 30. Total: 900 hojas

Número de flores[5]:

Tabla 4-2: Tamaño de muestra recomendado para cada variable de flor (codificación en la **Tabla 5-2**).

Variable	Número de flores observadas por clon	Número de flores observadas en total
Xf1, Xf2, Xf3 y Xf4	20	600
Xf5 y Xf6	15	450
Xf7, Xf8, Xf9, Xf10 y Xf11	10	300
Xf12	6	180
Xf13	5	150

Realizado por: Gabriela J. Obregón O. 2018

[5] Debido a la rapidez con la que se deteriora la flor una vez recolectada y a la limitación de tiempo, no fue posible medir todas las variables sobre una misma flor ni sobre todas las flores.

En el caso de los marcadores moleculares como muestra solo es necesario un pedazo de hoja por clon para la extracción de ADN.

2.6 Variables morfológicas

Las variables de caracterización morfológica consideradas fueron las consensuadas por los expertos del Programa Nacional de Cacao, donde se proporcionó una lista de las más importantes según lo recomendado por Bekele & Butler (BEKELE y BUTLER 2000) y, especialmente, por Soria y Enríquez (CENTRO AGRONÓMICO Y TROPICAL DE INVESTIGACIÓN Y ENSEÑANZA (CATIE) 1981) que identificaron las características más útiles para la descripción y un tamaño de muestra recomendado que toma en cuenta la variabilidad que se ha observado en cada una. Estas características están asociadas a los órganos que se sabe están menos influenciados por el medio ambiente: flor, mazorca, hojas y semillas. En la **Tabla 5-2** se describe cada una de las variables morfológicas.

Tabla 5-2: Descripción de las variables de caracterización morfológica.

Unidad de análisis	**Nombre de la variable**	**Codificación**	**Tipo y modalidades**	**Instrumento de medición**	**Unidad de medida**
Mazorca	Largo de la mazorca	X_{M1}	Cuantitativa: Intervalo de la recta real	Calibrador grande	cm
	Ancho de la mazorca	X_{M2}	Cuantitativa: Intervalo de la recta real	Calibrador grande	cm
	Peso de la mazorca	X_{M3}	Cuantitativa: Intervalo de la recta real	Balanza	g
	Peso de cáscara	X_{M4}	Cuantitativa: Intervalo de la recta real	Balanza	g
	Espesor de cáscara en el lomo	X_{M5}	Cuantitativa: Intervalo de la recta real	Calibrador pequeño	cm
	Espesor de cáscara en el surco	X_{M6}	Cuantitativa: Intervalo de la recta real	Calibrador pequeño	cm
	Número de semillas por mazorca	X_{M7}	Cuantitativa: Números naturales	Conteo manual	Unidad
	Peso total de semillas	X_{M8}	Cuantitativa: Intervalo de la recta real	Balanza	g
	Peso de 5 semillas húmedas con pulpa y testa	X_{M9}	Cuantitativa: Intervalo de la recta real	Balanza	g
	Peso de 5 semillas húmedas sin pulpa y testa	X_{M10}	Cuantitativa: Intervalo de la recta real	Balanza	g

	Peso de pulpa y testa de 5 semillas	X_{M11}	Cuantitativa: Intervalo de la recta real	$X_{M9} - X_{M10}$	g
	Peso de 5 semillas secas	X_{M12}	Cuantitativa: Intervalo de la recta real	Balanza	g
	Forma de mazorca	X_{M13}	Cuantitativa: Intervalo de la recta real	$\frac{X_{M1}}{X_{M2}}$	Cociente
Semilla	Largo de semilla	X_{S1}	Cuantitativa: Intervalo de la recta real	Calibrador pequeño	cm
	Ancho de semilla	X_{S2}	Cuantitativa: Intervalo de la recta real	Calibrador pequeño	cm
	Espesor de semilla	X_{S3}	Cuantitativa: Intervalo de la recta real	Calibrador pequeño	cm
	Color de semilla	X_{S4}	Cualitativa: Rojo Claro, Rojo Oscuro, Blanco	Observación	Tipo de color
Hoja	Largo de la hoja	X_{H1}	Cuantitativa: Intervalo de la recta real	Regla	cm
	Ancho de la hoja	X_{H2}	Cuantitativa: Intervalo de la recta real	Regla	cm
	Relación largo-ancho	X_{H3}	Cuantitativa: Intervalo de la recta real	$\frac{X_{H1}}{X_{H2}}$	Cociente
	Largo desde la base hasta el punto más ancho del limbo	X_{H4}	Cuantitativa: Intervalo de la recta real	Regla	cm
	Forma de la hoja (cociente)	$X_{H5.1}$	Cuantitativa: Intervalo de la recta real	$\frac{X_{H1}}{X_{H4}}$	Cociente
	Ángulo basal	X_{H6}	Cuantitativa: $[0°; 180°]$	Graduador	Grados
	Ángulo apical	X_{H7}	Cuantitativa: $[0°; 180°]$	Graduador	Grados
	Forma de la hoja (categoría)	$X_{H5.2}$	Cualitativa: Ovalada, Oblonga, Elíptica	$X_{H5} < 2$: ovalada $X_{H5} = 2$: oblonga $X_{H5} > 2$: elíptica	Tipo de forma
Flor	Pigmentación del pedúnculo	X_{F1}	Cualitativa: Pigmentado, Medianamente Pigmentado, No pigmentado	Observación	Tipo de pigmentación
	Pigmentación de sépalos	X_{F2}	Cualitativa: Pigmentado, Medianamente	Observación	Tipo de pigmentación

			Pigmentado, No pigmentado		
	Pigmentación de filamentos estaminales	X_{F3}	Cualitativa: Pigmentado, Medianamente Pigmentado, No pigmentado	Observación	Tipo de pigmentación
	Ancho de un sépalo	X_{F4}	Cuantitativa: Intervalo de la recta real	Papel milimetrado	mm
	Largo de una lígula	X_{F5}	Cuantitativa: Intervalo de la recta real	Papel milimetrado	mm
	Ancho de una lígula	X_{F6}	Cuantitativa: Intervalo de la recta real	Papel milimetrado	mm
	Largo de un sépalo	X_{F7}	Cuantitativa: Intervalo de la recta real	Papel milimetrado	mm
	Largo del estilo	X_{F8}	Cuantitativa: Intervalo de la recta real	Papel milimetrado	mm
	Largo del ovario	X_{F9}	Cuantitativa: Intervalo de la recta real	Papel milimetrado	mm
	Ancho del ovario	X_{F10}	Cuantitativa: Intervalo de la recta real	Papel milimetrado	mm
	Largo del pedúnculo	X_{F11}	Cuantitativa: Intervalo de la recta real	Papel milimetrado	mm
	Largo de un estaminoide	X_{F12}	Cuantitativa: Intervalo de la recta real	Papel milimetrado	mm
	Número de óvulos en el ovario	X_{F13}	Cuantitativa: Números Naturales	Conteo en estereoscopio	Unidad

Realizado por: Gabriela J. Obregón O. 2018

2.6.1 Selección de muestra y técnicas de recolección de datos morfológicos

Los datos de los 26 clones de tipo Criollo fueron medidos durante el año 2014 (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014) y proporcionados por dicho autor, mientras que los datos de los 4 clones testigo fueron tomados por la autora del 5 al 16 de marzo de 2018, siguiendo el mismo procedimiento, luego de una capacitación. Las mazorcas, hojas y flores de todos los clones, excepto del CCN-51, fueron recolectados en los jardines clonales de la Estación Experimental Litoral Sur (EELS). Los materiales del clon CCN-51 se recolectaron en una finca cercana. Todo el proceso de medición se llevó a cabo en el laboratorio del Programa de Cacao de la EELS.

2.6.1.1 Procedimiento de recolección y medición de mazorcas y semillas

1. Preparación de materiales para recolección: fundas plásticas, tijeras para podar, rotuladores, etiquetas para árbol.
2. Visita a jardín clonal: toma de muestra de frutos maduros y sanos (20 unidades por clon) de diferentes plantas de un mismo clon, etiquetada de árbol y fruto.

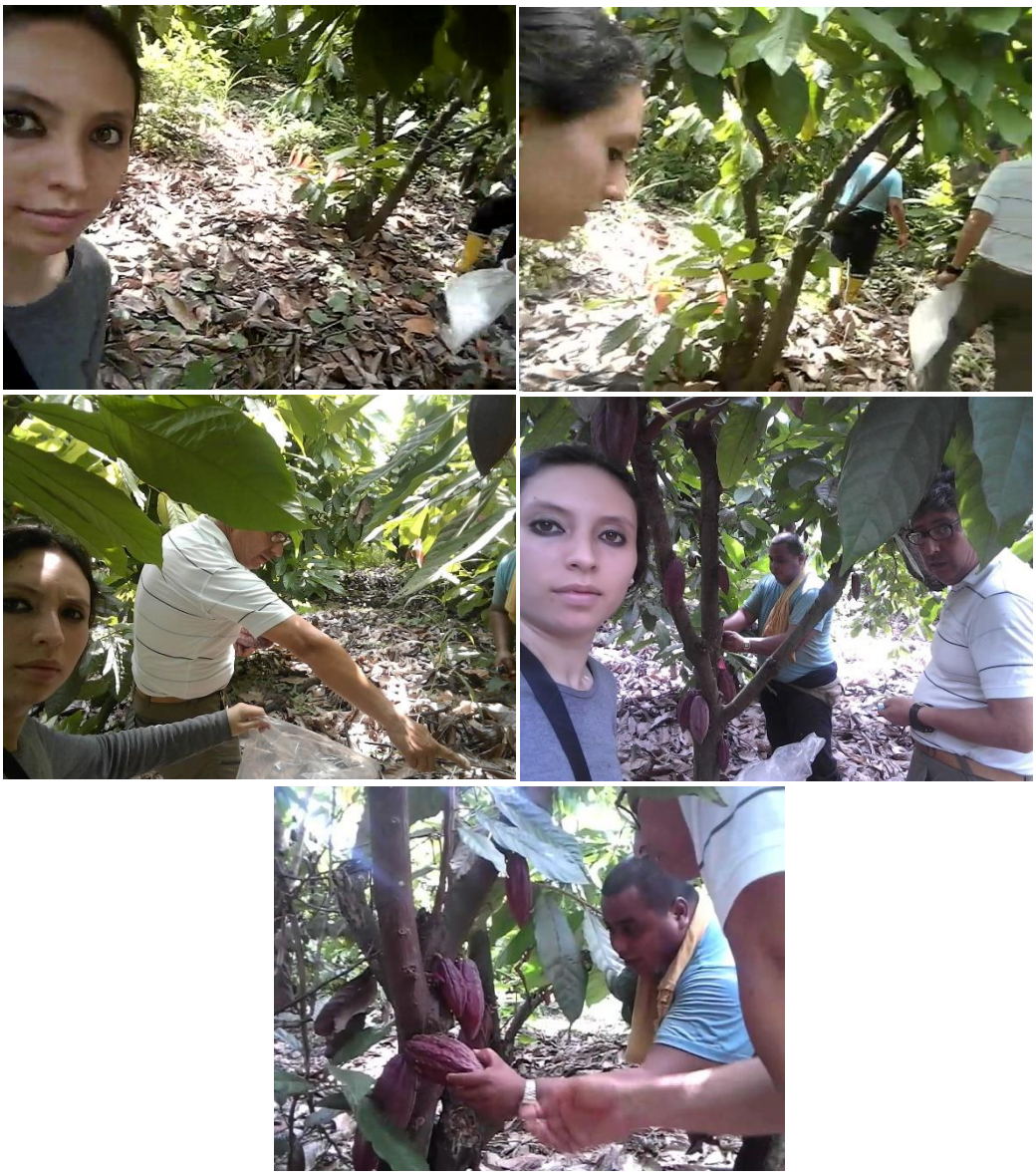

Gráfico 1-2: Visita al jardín clonal para recolección de mazorcas.

3. Una vez en el laboratorio, etiquetado de cajas Petri para almacenar las semillas.
4. Preparación de herramientas e instrumentos a utilizar: machete, cuchillo, calibradores vernier, franelas, balanza de precisión electrónica, bandejas de plástico y horno para secado.

5. Medición de mazorcas: Medición de Xm1, Xm2, Xm13, Xm3, Xm4, Xm5 y Xm6. Partición de la mazorca, conteo para Xm7, y Xm8. Selección de 5 semillas al azar para Xm9, Xm10 y Xm11.
6. Medición de Xs1, Xs2, Xs3 y Xs4 en cada una de las 5 semillas seleccionadas.
7. Medición de Xm12 ubicando las semillas en cajas Petri etiquetadas. Fueron secadas durante 24 horas en un horno a 60ºC.
8. La matriz de datos de mazorcas es de tamaño 600×13. La matriz de datos de semillas es de tamaño 3000×3 (más una variable cualitativa).

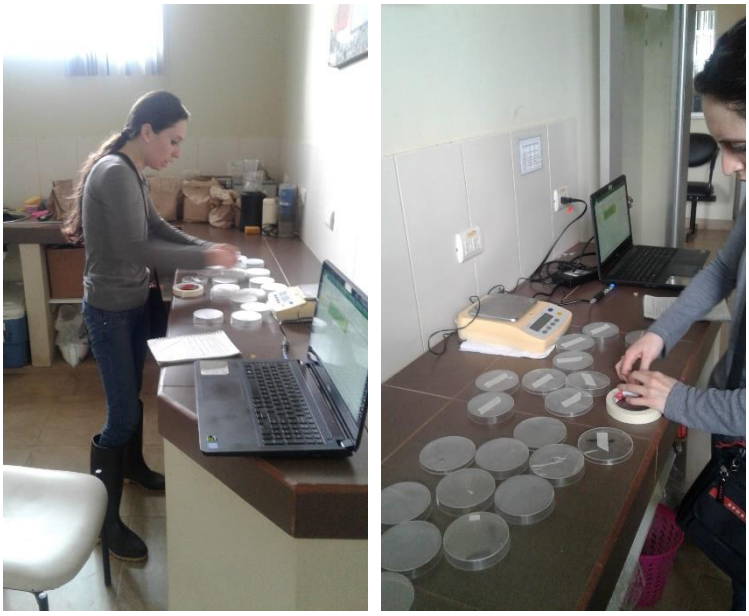

Gráfico 2-2: Etiquetado de cajas Petri para almacenar semillas por mazorca.

Gráfico 3-2: Medición de diámetro y peso de una mazorca.

Gráfico 4-2: Medición de peso de cáscara, espesor en lomo y surco, conteo y peso de semillas.

Gráfico 5-2: Medición y observación de color de 5 semillas por mazorca seleccionadas al azar.

Gráfico 6-2: Colocación de semillas en el horno para obtener el peso seco.

2.6.1.2 Procedimiento de recolección y medición de hojas

1. Preparación de materiales para recolección de hojas: fundas plásticas, rotuladores, tijeras para podar.

2. Visita a jardín clonal: toma de muestra de hojas (30 unidades) en buen estado, sin insectos, jóvenes, color verde y completas. Se selecciona la segunda o tercera de una rama, de diferentes plantas de un mismo clon.
3. Una vez en laboratorio, se colocan en refrigeración a fin mantenerlas frescas y en buen estado.
4. Preparación de material necesario para medir y determinar formas: pliegos de papel periódico, papel de calcar, reglas, graduador, papel bond, lápiz y borrador.
5. Calcado de cada hoja desde la base hasta el ápice.
6. Medición de cada hoja calcada: Xh1, Xh2, Xh3, Xh4, Xh5.1, Xh6 y Xh7.

La matriz de datos de hojas es de tamaño 900×7 (más una variable cualitativa).

Gráfico 7-2: Recolección de hojas en el campo.

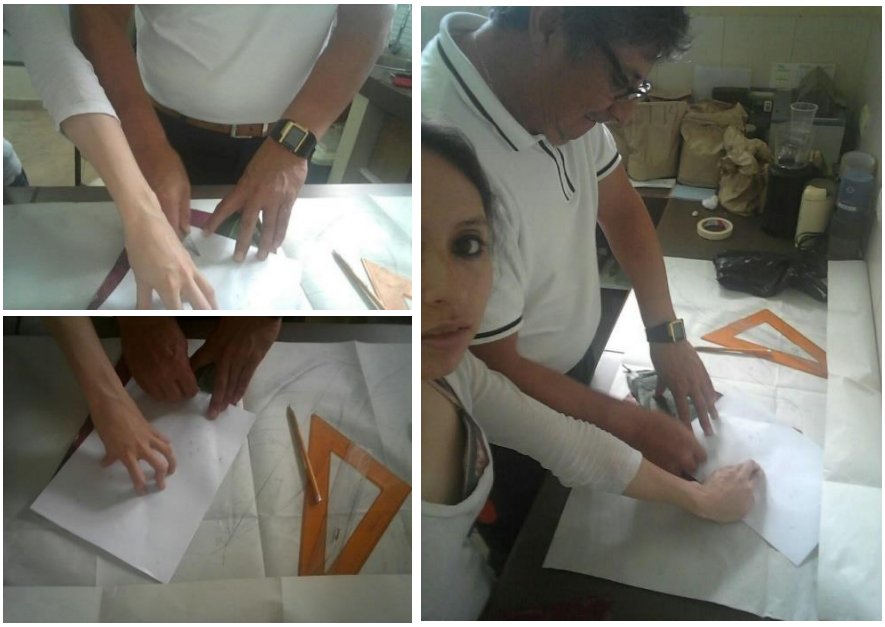

Gráfico 8-2: Calcado de hojas desde la base hasta el ápice en papel periódico.

Gráfico 9-2: Medición de variables de hoja en el papel periódico.

2.6.1.3 Procedimiento de recolección y medición de flores

1. Etiquetado y preparación de cajas Petri para almacenamiento del material.
2. Visita a jardín clonal: toma de muestras de flores (20 unidades) en buen estado, de ser posible de varias plantas de un mismo clon.

3. Una vez en el laboratorio colocar muestras en refrigeración o con algodón humedecido para retrasar la oxidación.
4. En cada flor observación de Xf1, Xf2 y Xf3.
5. Medición de las diferentes partes de la flor (Xf4-Xf12) con papel milimetrado e instrumentos de manipulación adecuados (pinzas, lupa, estilete, lámpara).
6. Conteo de Xf13 mediante estereoscopio.

Gráfico 10-2: Medición de variables de flor directamente observables.

Gráfico 11-2: Observación de ovario en estereoscopio para conteo de número de óvulos.

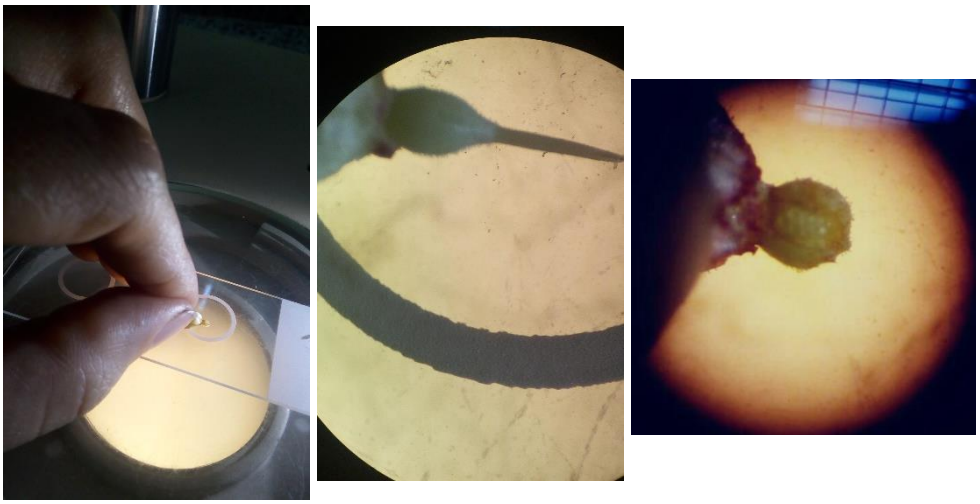

Gráfico 12-2: Vista del ovario en el estereoscopio antes y después de pelar un surco para conteo de óvulos.

2.7 Variables moleculares

Se utilizaron 20 marcadores moleculares Microsatélites SSR. Éstos fueron seleccionados por los técnicos del laboratorio de Biotecnología del INIAP Santa Catalina según lo recomendado internacionalmente, y son:

mTcCIR1, mTcCIR6, mTcCIR7, mTcCIR8, mTcCIR10, mTcCIR11, mTcCIR12, mTcCIR15, mTcCIR22, mTcCIR26, mTcCIR29, mTcCIR33, mTcCIR37, mTcCIR40, mTcCIR58, mTcCIR60, mTcCIR230, mTcCIR274, mTcCIR290, mTcCIR293

2.7.1 Selección de muestra y técnicas de recolección de datos moleculares

En el *Anexo C* se muestra el protocolo que se siguió para la toma de muestras foliares en campo. Estas muestras fueron trasladadas a la Estación Experimental Santa Catalina del INIAP para la extracción del ADN y posterior obtención de los datos de marcadores moleculares. Se realizó el genotipaje de 20 marcadores Microsatélites con la tecnología M13 Tailing (MORILLO VELASTEGUI y MIÑO CASTRO 2011). Todo se llevó a cabo en el laboratorio del Departamento Nacional de Biotecnología del INIAP.

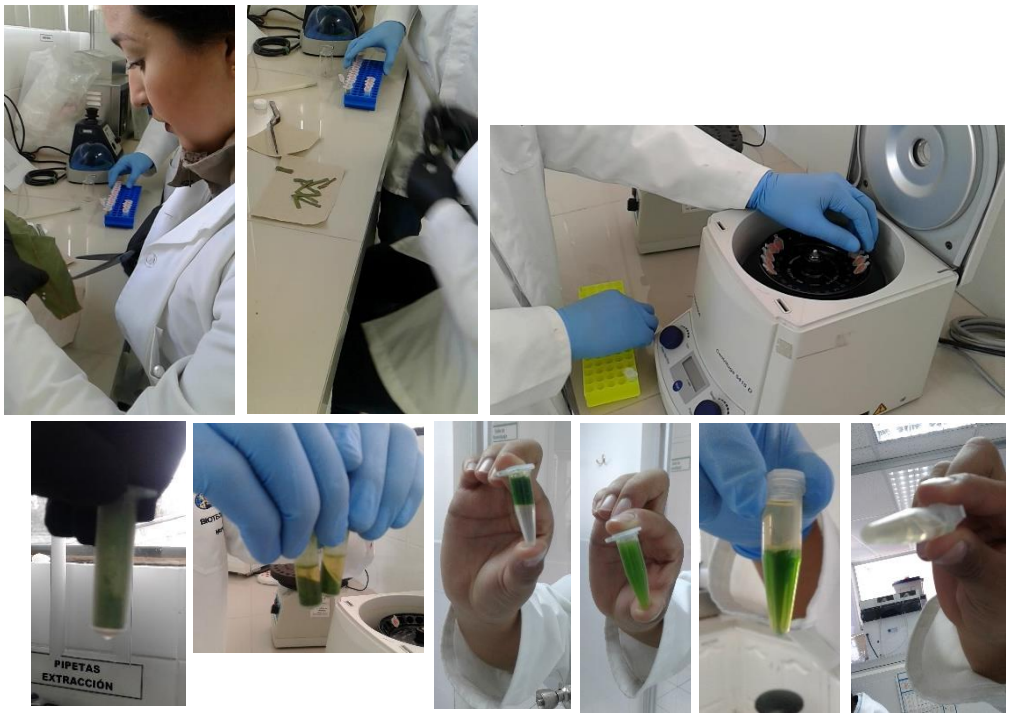

Gráfico 13-2: Diferentes fases para la extracción de ADN.

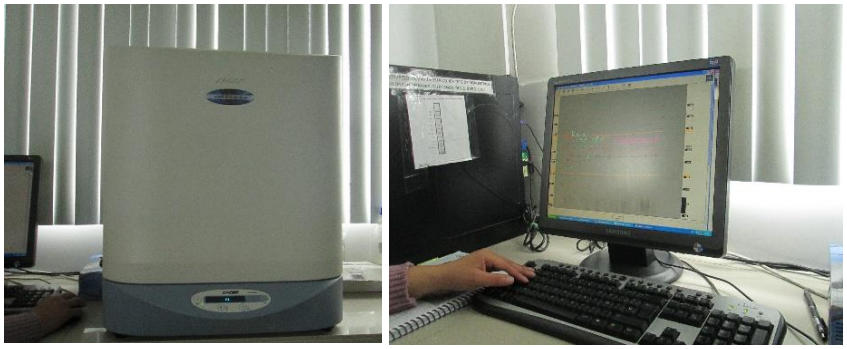

Gráfico 14-2: Obtención de los datos de marcadores moleculares con el equipo LI-COR 4300.

Tabla 6-2: Ejemplo de datos de caracterización molecular.

Primer		**mTcCIR230**	
No.	Código muestra	**ALELO 1**	**ALELO 2**
2	C2	254	254
3	C3	248	248
4	C4	248	254
16	C16	236	254
23	C23	234	234
26	C26	246	248
28	EET-111	248	252
30	CCN-51	234	252

Fuente: Matriz de datos moleculares.

La matriz de datos moleculares[6] está formada por 30 filas (clones), y 20 marcadores moleculares en las columnas. Cada marcador molecular presenta 2 posibilidades de alelos, cada uno con un valor que corresponde a los pares de bases nitrogenadas presentes en el locus. La matriz inicial es de orden 30×20. A modo de ejemplo en la **Tabla 6-2** se extraen de esta matriz los valores obtenidos para el marcador molecular mTcCIR230 en los clones C2, C3, C4, C16, C23, C26, EET-111 y CCN-51.

En el caso de este marcador molecular se muestran todos los valores o categorías que se presentaron en cada alelo en la **Tabla 7-2**.

Tabla 7-2: Valores que aparecieron en cada alelo para el marcador mTcCIR230.

mTcCIR230	
ALELO 1	**ALELO 2**
234	234
236	248
246	252
248	254
254	

Así el número de bandas diferentes observadas en el locus correspondiente al marcador mTcCIR230 es de 6: 234, 236, 246, 248, 252, 254.

2.8 Análisis estadístico

Para el análisis de datos de caracterización morfológica el Instituto Internacional de Recursos Fitogenéticos (IPGRI) recomienda varios métodos para obtener conclusiones acerca de la variabilidad y la utilidad del germoplasma. Se recomienda el uso de la media aritmética, el rango, la desviación estándar y el coeficiente de variación para un análisis exploratorio de datos cuantitativos, tablas de frecuencias para datos cualitativos, medidas de similitud como el índice de similitud, distancias y coeficientes de correlación, según sea el tipo de dato (INTERNATIONAL PLANT GENETIC RESOURCES INSTITUTE (IPGRI) 2003), y técnicas de análisis multivariado de ordenación (Análisis de Componentes Principales, Análisis de Coordenadas Principales, Análisis Factorial, etc.) o de clasificación (Análisis Cluster, Método de Ward, Árbol de Mínima Distancia, etc.) según sea el caso (BRAMARDI, y otros 2005).

En la caracterización morfológica desarrollada por Pesántez (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014) se desarrollaron los siguientes análisis: Estadística descriptiva, Análisis de Componentes Principales, Análisis Cluster que generó un

[6] En el *Anexo E* se adjunta la matriz, así como la foto-documentación de la cual provienen los datos en el *Anexo D*.

dendrograma con el Método de Ward mediante coeficiente de similaridad general de Gower, prueba Chi Cuadrado para variables cualitativas, y prueba de Duncan para comparación de medias en variables cuantitativas, para determinar las variables que más discriminan los clones.

Para datos de caracterización molecular Bruno y Balzarini (BRUNO y BALZARINI 2010) recomiendan calcular índices de similitud para realizar un Análisis de Coordenadas Principales. Martínez (1995) citado por Quiroz (QUIROZ VERA 2002) establece que si el propósito es evaluar la variabilidad, clasificación, estructura y composición genética de las poblaciones, los métodos estadísticos más utilizados son distancias genéticas, índices de similitud , dendrogramas y Coordenadas Principales.

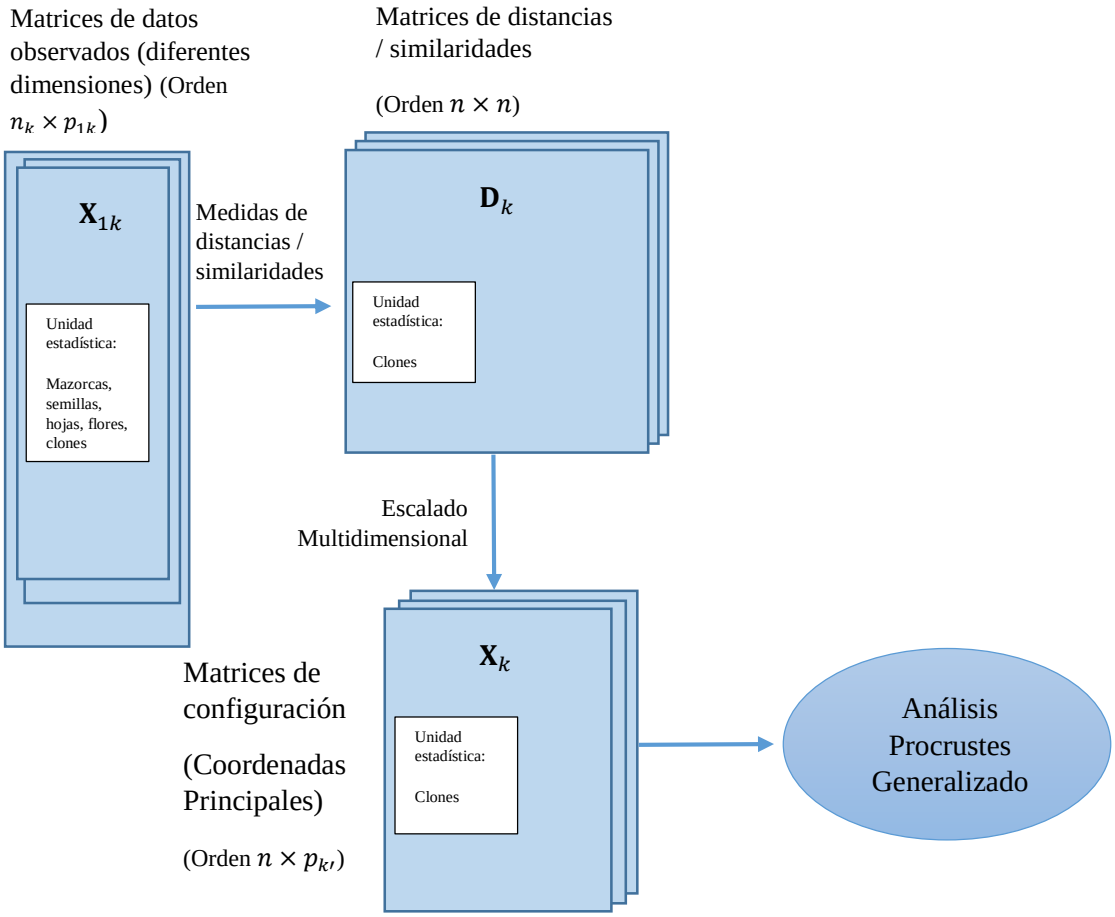

Gráfico 15-2: Esquema del pre-proceso de análisis. $k = 1, 2, \ldots, m$.

Realizado por: Gabriela J. Obregón O. 2018.

Para el tratamiento conjunto de marcadores, como ya se mencionó anteriormente, los autores Bramardi et al. (BRAMARDI, y otros 2005) y Bruno y Balzarini (BRAMARDI, y otros 2005), entre otras técnicas complementarias (como el test de Mantel), recomiendan el Análisis Procrustes Generalizado (APG). El Escalado Multidimensional (u otros tipos de métodos multivariantes según los datos que se tengan) se recomienda para producir configuraciones limpias previas a un

Análisis Procrustes (GOWER y DIJKSTERHUIS, Procrustes Problems 2004), lo cual, para el caso del presente estudio, se ilustra en el **Gráfico 15-2**, como un pre-proceso de análisis basado en el esquema de, y según lo recomendado por, Gower y Dijksterhuis (GOWER y DIJKSTERHUIS, Procrustes Problems 2004).

Se ilustra la representación gráfica del proceso de análisis en el **Gráfico 16-2**, donde se lo divide en las fases de: preparación, pre-proceso, input, proceso (Análisis Procrustes Generalizado), output y post proceso.

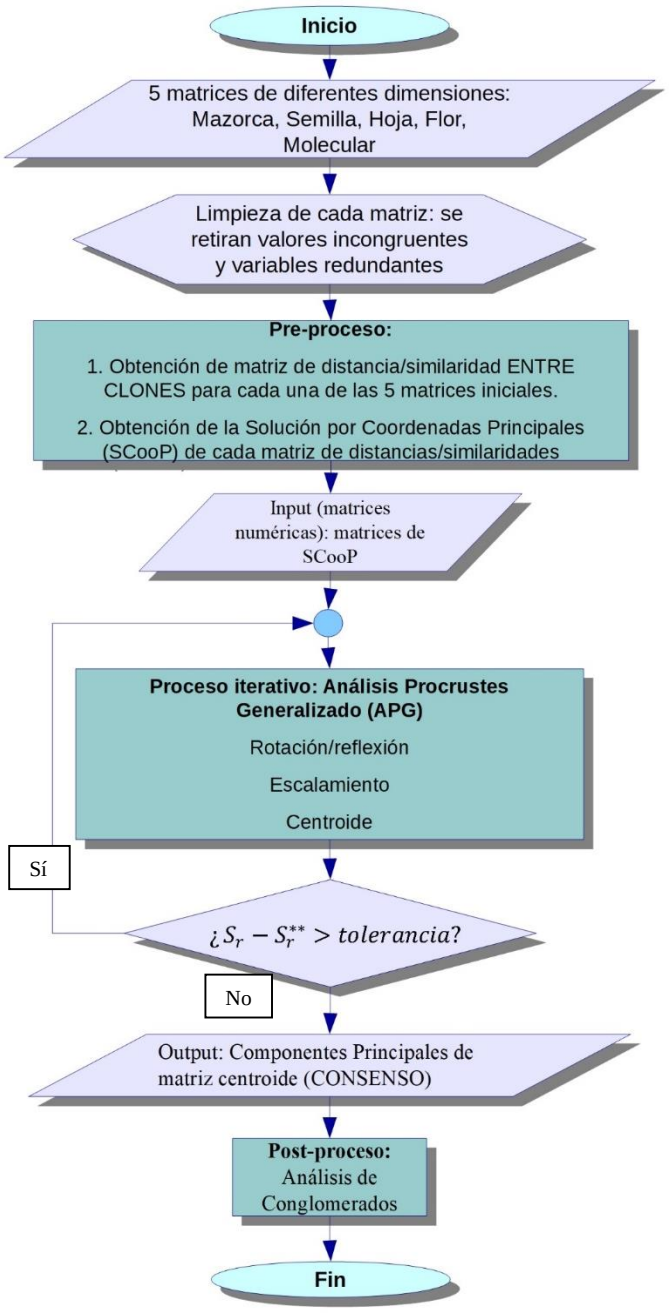

Gráfico 16-2: Representación gráfica del proceso de análisis.[7]

Realizado por: Gabriela J. Obregón O. 2018.

2.8.1 Limpieza de datos

[7] En este caso para el APG no se considera traslación porque no hace falta, ya que las coordenadas principales ya están centradas en el origen.

El paso inicial en el procesamiento de datos fue revisar que todos los datos estén dentro del campo de variación posible para cada variable. Se eliminaron las filas que presentaron datos con valores extremos incongruentes, así como con datos faltantes.

2.8.2 Análisis descriptivo univariado

El análisis estadístico descriptivo se realizó para explorar las variables en estudio (Software: R) cuyo código se adjunta en la nube (https://onedrive.live.com/?authkey=%21AIJE2OcZQnWrOgE&id=CA77A5953C07ED0C%21140440&cid=CA77A5953C07ED0C). Se realizó un análisis estadístico-descriptivo univariado con la totalidad de los datos originales para tener un panorama general de la distribución mediante histogramas y diagramas de caja para variables cuantitativas y diagramas de pastel para visualizar la proporción de las categorías de variables cualitativas, así como de los indicadores de cada variable. Se calcularon valores mínimo y máximo, rango, media, mediana, moda, cuartiles, varianza, desviación estándar, coeficiente de variación, coeficiente de asimetría e indicador de exceso de kurtosis.

2.8.3 Pruebas no paramétricas.

Se realizaron pruebas no paramétricas para determinar si hay diferencia significativa entre los clones dentro de cada variable morfológica utilizando todos los valores observados disponibles para cada variable (Software: R). Se aplicaron tests análogos no paramétricos al ANOVA de una vía[8]: el test de Kruskal-Wallis del paquete 'stats' y el test de Análisis de Varianza de una vía basado en Rango (HETTMANSPERGER y MCKEAN 2011) del paquete 'Rfit'.

Las hipótesis planteadas fueron:

$$H_0: Las\ medianas\ de\ los\ clones\ de\ la\ variable\ en\ cuestión\ son\ iguales$$

$$H_1: Al\ menos\ uno\ de\ los\ clones\ tiene\ una\ mediana\ distinta\ a\ la\ de\ los\ otros$$

2.8.4 Variables redundantes

Con base en el teorema de la dimensión se determinaron el número de variables linealmente independientes, y se retiraron de la matriz las variables que se pueden expresar como combinación

[8] Debido a que la mayoría de variables no cumplió con varios supuestos que validan un ANOVA.

lineal de las demás, para lo cual se observa el número de autovalores nulos de la matriz de covarianzas. Hay tantas variables redundantes como autovalores nulos.

2.8.5 AMOVA

Se realizó un Análisis de varianza molecular (AMOVA) en el Software GenAlex. Se calculó el AMOVA con los siguientes parámetros: 20 loci (marcadores moleculares), 30 muestras (clones), 3 poblaciones (Criollos, Testigos de origen nacional y Testigos de origen exterior), cada una con tamaño de muestra 26, 2 y 2 respectivamente.

2.8.6 Análisis descriptivo multivariado

Se analizaron las matrices de covarianzas y de correlaciones y se calcularon medidas globales de variabilidad.

2.8.6.1 Matriz de flores

Para que la unidad de análisis sea cada clon, se estandarizaron de forma univariada cada una de las variables cuantitativas de flor, y, debido a que no todas las variables pudieron ser medidas sobre todas las flores, se colocó como dato para cada clon el promedio por clon. La matriz de datos de flor es de orden 30×13. Se calculó la distancia de Mahalanobis entre clones.

Variables utilizadas:

Xf1.1: Promedio de ancho de sépalo en 20 flores observadas

Xf2.1: Promedio de largo de lígula en 15 flores observadas

Xf3.1: Promedio de ancho de lígula en 15 flores observadas

Xf4.1: Promedio de largo de sépalo en 10 flores observadas

Xf5.1: Promedio de largo del estilo en 10 flores observadas

Xf6.1: Promedio de largo del ovario en 10 flores observadas

Xf7.1: Promedio de ancho del ovario en 10 flores observadas

Xf8.1: Promedio de largo del pedúnculo en 10 flores observadas

Xf9.1: Promedio de largo del estaminoide en 6 flores observadas

Xf10.1: Promedio de óvulos observados en el ovario de 5 flores

2.8.7 Análisis de Coordenadas Principales

Las matrices de datos morfológicos observados donde las unidades de análisis fueron mazorcas, semillas, hojas y flores, fueron el punto de partida. Se incorporaron las variables cualitativas Color de semilla, Pigmentación de pedúnculo floral, Pigmentación de sépalos y Pigmentación de filamentos estaminales[9], que, por su naturaleza de alguna manera ordinal de intensidad de pigmentación, se asignaron valores a cada categoría, de la siguiente forma:

Xs4: 0.0 – Blanco, 0.5 – Rojo claro, 1.0 – Rojo oscuro

Xf1, Xf2 y Xf3: 0.0 – No pigmentado, 0.5 – Medianamente pigmentado, 1.0 – Pigmentado

Se aplicó estandarización univariada a cada matriz morfológica debido a que las intensidades totales y varianzas de sus variables fueron muy diferentes. Cada una de las matrices morfológicas originales es de diferente orden debido a que la unidad de análisis de cada una (mazorcas, semillas, hojas, flores) y su numerosidad es diferente, así como también las variables observadas son diferentes para cada una. Como el objeto de interés no son en sí las unidades de análisis mencionadas, sino más bien los clones de cacao, lo que se requiere es una matriz que tenga a cada clon como unidad de análisis. Para ello se calcularon las distancias entre poblaciones con la fórmula basada en la distancia de Mahalanobis para obtener la matriz de distancias, constituyendo cada clon una población (Software: se desarrolló un código en R para automatizar el proceso, (*Apéndice A*)).

Se aplicó Escalado Multidimensional a cada matriz de distancias siguiendo el procedimiento detallado en marco teórico (Software: Excel 2016). Se obtuvieron en total 5 configuraciones diferentes conformadas por la solución por coordenadas principales obtenida con escalado multidimensional métrico y no métrico según el caso. Las 5 configuraciones se denominaron: Molecular, Mazorca, Semilla, Hoja y Flor.

Se adjuntan las matrices de distancia y de similaridad de las que se partió para el Análisis de Coordenadas Principales en la *nube* (link: https://onedrive.live.com/?authkey=%21AIJE2OcZQnWrOgE&id=CA77A5953C07ED0C%21140440&cid=CA77A5953C07ED0C).

[9] No se incluyó la variable cualitativa Forma de la hoja, que se categorizó a partir de operaciones entre otras variables, debido a que la variable cuantitativa de la que partió, fue redundante.

2.8.7.1 Configuración Molecular

Para la matriz de datos moleculares desde un inicio la unidad de análisis fueron los clones de cacao. Para el presente estudio tenemos que el cacao es un organismo diploide, es decir, que para un individuo dentro de un mismo locus (marcador molecular determinado) hay dos posibilidades de alelos, pudiendo éstos ser los mismos (estado homocigoto) o pudiendo ser diferentes (estado heterocigoto). Además, tenemos que los marcadores moleculares SSR son codominantes, esto es, que permiten distinguir patrones de bandas entre homocigotos y heterocigotos. Siendo así, se calculó el índice de similaridad propuesto por Kosman y Leonard para obtener la matriz de similaridades (Software: Excel 2016).

2.8.8 Análisis Procrustes Generalizado

Se realizó un APG ingresando las configuraciones de coordenadas principales en el programa Infostat 2018[10]. El procedimiento en Infostat se presenta a continuación:

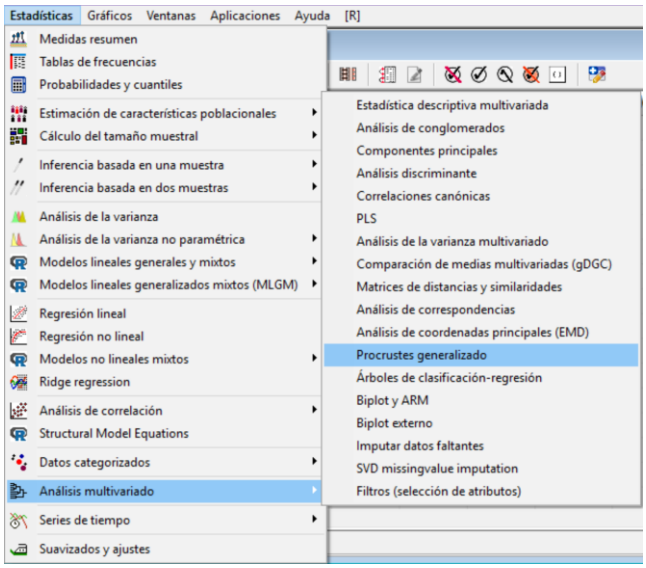

Gráfico 17-2: Análisis Procrustes Generalizado en Infostat 2018.

Una vez ingresados los datos con los clones en las filas y las coordenadas principales en las columnas, se siguen los siguientes pasos mostrados en el **Gráfico 17-2**.

En la sección ‘Variables’ se ubican todas las variables a consensuar, en este caso son las coordenadas principales, y clic en ‘Aceptar’:

[10] Adicionalmente se analizaron en R con el paquete *FactoMineR* para comparar y corroborar lo obtenido, donde los resultados fueron prácticamente iguales (no exactamente iguales debido a la diferencia entre el número de iteraciones y el valor umbral establecido), cuyo código se adjunta en el *Apéndice B*.

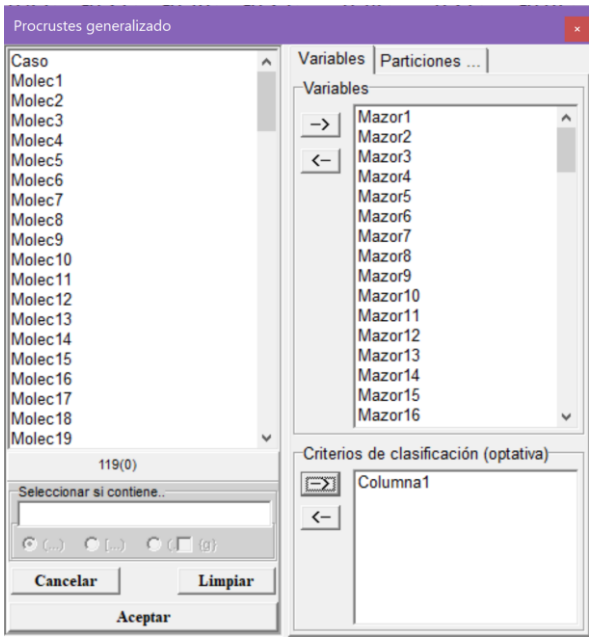

Gráfico 18-2: Selección de variables a consensuar y criterio de clasificación.

Se seleccionan todas las variables pertenecientes a la primera configuración y se da clic derecho y 'nuevo'. Se repite el proceso formando un grupo para cada configuración. A continuación se seleccionan las opciones deseadas como 'Guardar' para almacenar el número de coordenadas deseado del consenso resultante, así como la opción de 'Graficar' y 'Escalar'. Finalmente clic en 'Aceptar'.

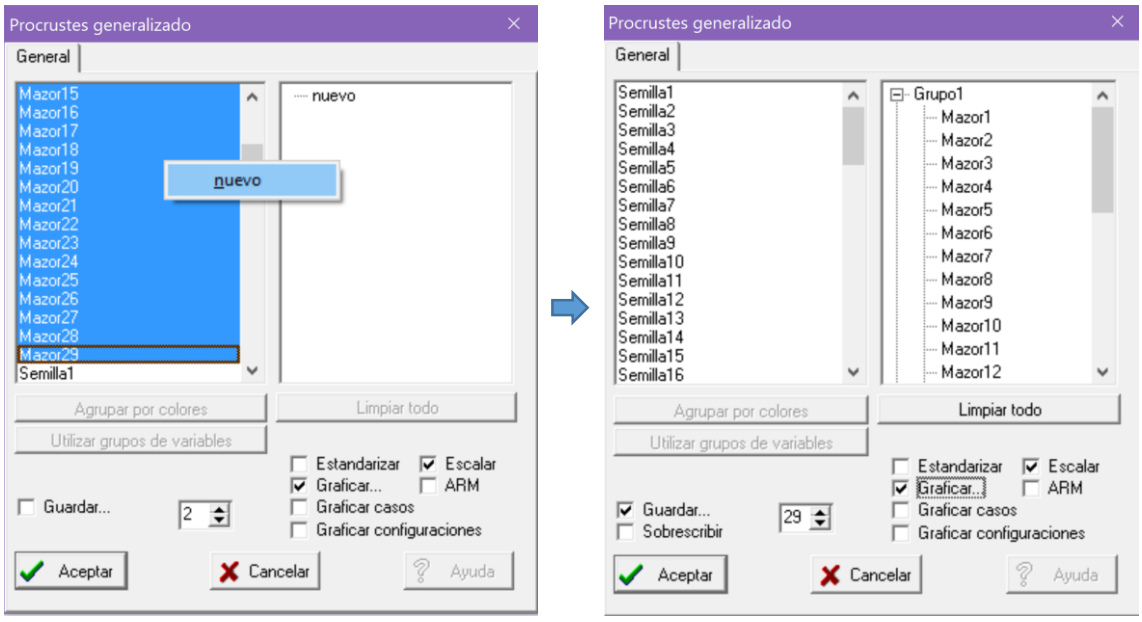

Gráfico 19-2: Determinación de configuraciones (grupos) a consensuar y configuración de opciones de resultados.

Al obtener un consenso con las 5 configuraciones simultáneamente se notó que se estaba dando mayor ponderación al aspecto morfológico debido a que de éste se tenían 4 configuraciones,

mientras que sólo se tenía 1 configuración molecular, y lo deseado era que lo morfológico y lo molecular tengan el mismo peso. Por este motivo se decidió obtener un consenso sólo entre configuraciones de caracterización morfológica, y finalmente otro entre éste y la configuración molecular. De las configuraciones consenso el programa Infostat entregó componentes principales, mismas que fueron el punto de partida para posteriormente obtener dendrogramas con agrupamiento jerárquico de Ward para ilustrar las agrupaciones.

2.9 Post-proceso: Análisis de conglomerados

Se obtuvieron dendrogramas por agrupamiento jerárquico de Ward para cada configuración individual (coordenadas principales) y para cada configuración consenso (componentes principales del consenso) para determinar agrupaciones. Todos con base en la distancia euclídea, la cual se consideró más apropiada para "variables" de solución por coordenadas principales y solución por componentes principales debido a su naturaleza numérica continua, además de que están incorreladas. Se debe recalcar que el análisis de conglomerados se emplea en análisis exploratorio de datos, por lo que no es una técnica inferencial, sin embargo, permite determinar grupos homogéneos de clones con base en las variables que se midieron (VILLA MOYOTA 2012). Se realizó en Infostat 2018, cuyos pasos se presentan a continuación:

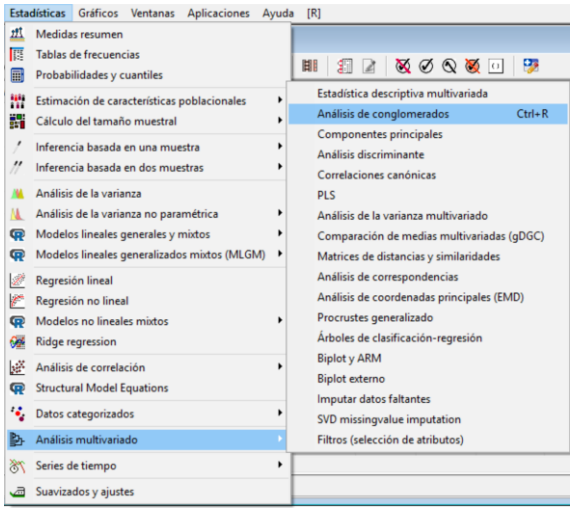

Gráfico 20-2: Pasos para realizar análisis de conglomerados en Infostat 2018.

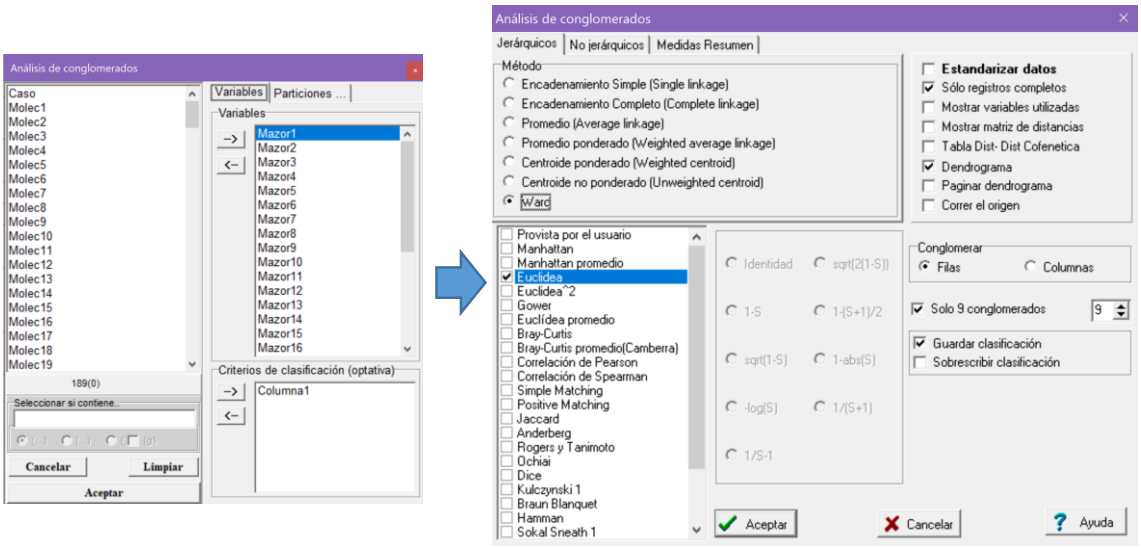

Gráfico 21-2: Selección de variables, método, distancia o similaridad, número de conglomerados y otras opciones.

CAPÍTULO III

3 RESULTADOS Y DISCUSIÓN

En el **Gráfico 1-3** se muestra la relación entre índice de mazorca (mazorcas/kg de cacao seco) y el rendimiento (kg/Ha/Año) de cada uno de los clones en estudio, en comparación con varios clones testigo (datos obtenidos por Pesántez (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014)), donde se aprecia cómo los clones testigo, que no son de tipo Criollo, presentaron un rendimiento mucho mayor, conformando un grupo que se distingue claramente. Se aprecia además cómo los 6 clones de tipo Criollo de mayor rendimiento conforman otro grupo, que destaca del resto de Criollos. Por tanto, se consideró de interés el seguir la pista de estos clones en los resultados presentados más adelante.

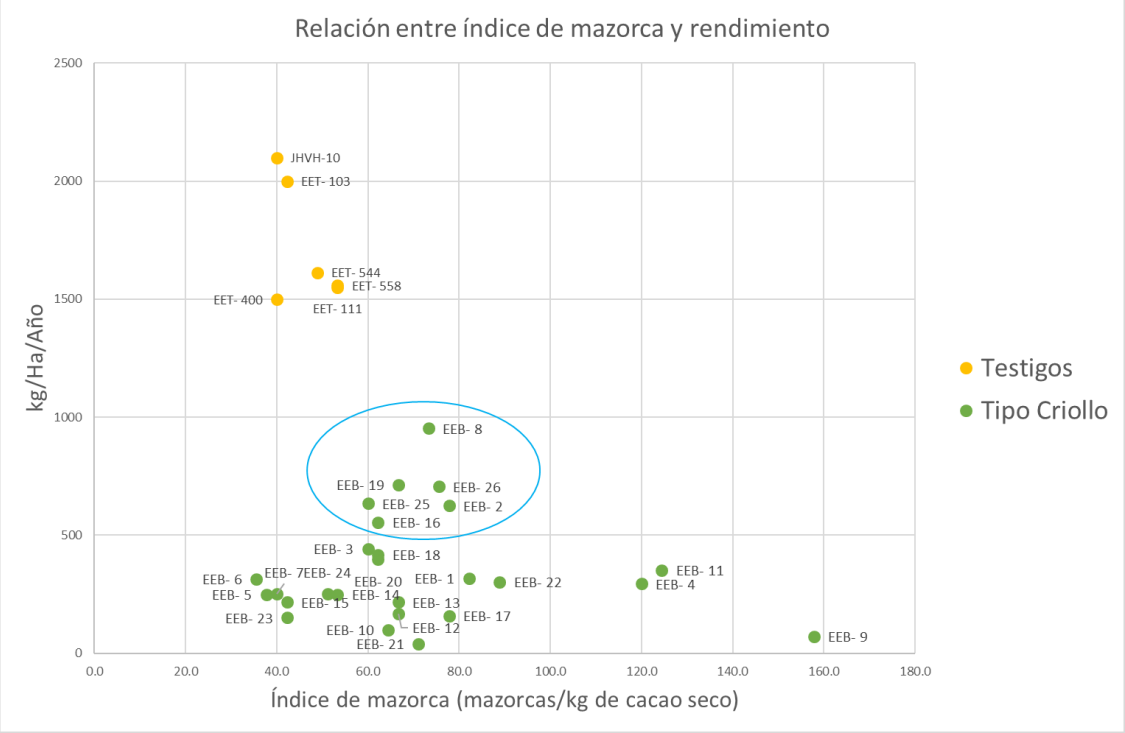

Gráfico 1-3: Clones de tipo Criollo que destacan en rendimiento e índice de mazorca.

Realizado por: Gabriela J. Obregón O. 2018.

3.1 Análisis descriptivo univariado inicial

Los resultados del análisis descriptivo univariado se muestran en tablas y gráficas en la nube (https://onedrive.live.com/?authkey=%21AIJE2OcZQnWrOgE&id=CA77A5953C07ED0C%21140440&cid=CA77A5953C07ED0C). Se comenta acerca de los valores del coeficiente de variación, pues según el IPGRI (INTERNATIONAL PLANT GENETIC RESOURCES INSTITUTE (IPGRI) 2003), si una variable presenta un coeficiente de variación mayor al 50% es porque tiene muy alta variabilidad en la especie, en cambio, si es menor al 20% indica que la especie puede tener poca variabilidad en ese carácter. En la **Tabla 1-3** se presentan los coeficientes de variación de cada variable que fueron obtenidos de forma univariada a partir de la totalidad de datos recolectados, y se compara con los coeficientes de variación obtenidos por Pesántez (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014), que fueron calculados considerando únicamente las medias de cada variable para cada uno de los clones en estudio.

Tabla 1-3: Coeficiente de Variación por variable.

Variable	Coeficiente de Variación obtenido de todos los datos	Coeficiente de Variación obtenido de los promedios por clon
Xm8	50%	26%
Xm4	39%	23%
Xm7	38%	26%
Xm3	38%	22%
Xm11	35%	*19%
Xf9	28%	19%
Xm9	27%	*17%
Xm6	27%	21%
Xm10	27%	*18%
Xm12	25%	*16%
Xf4	23%	14%
Xf11	22%	14%
Xh7	22%	11%
Xh4	21%	13%
Xm5	20%	16%
Xs3	19%	10%
Xh2	19%	17%
Xf8	19%	13%
Xf6	19%	13%
Xh1	18%	12%
Xm1	16%	16%
Xf10	15%	7%
Xf12	15%	10%
Xs2	14%	8%
Xf13	13%	12%
Xh3	13%	9%
Xf5	13%	8%
Xm13	13%	No disponible
Xm2	13%	14%
Xf7	12%	9%
Xs1	11%	6%
Xh5.1	10%	No disponible
Xh6	9%	12%

*Los datos para calcular este coeficiente de variación no fueron de 5 semillas, sino del peso promedio, es decir, con respecto a 1 semilla.

Codificación de color:

Unidad de análisis:	Mazorca	Flor	Hoja	Semilla

Escala de color del máximo al mínimo	Máximo	Mínimo
A partir de todos los datos	0.50	0.09
A partir de promedios por clon	0.26	0.06

Realizado por: Gabriela J. Obregón O. 2018

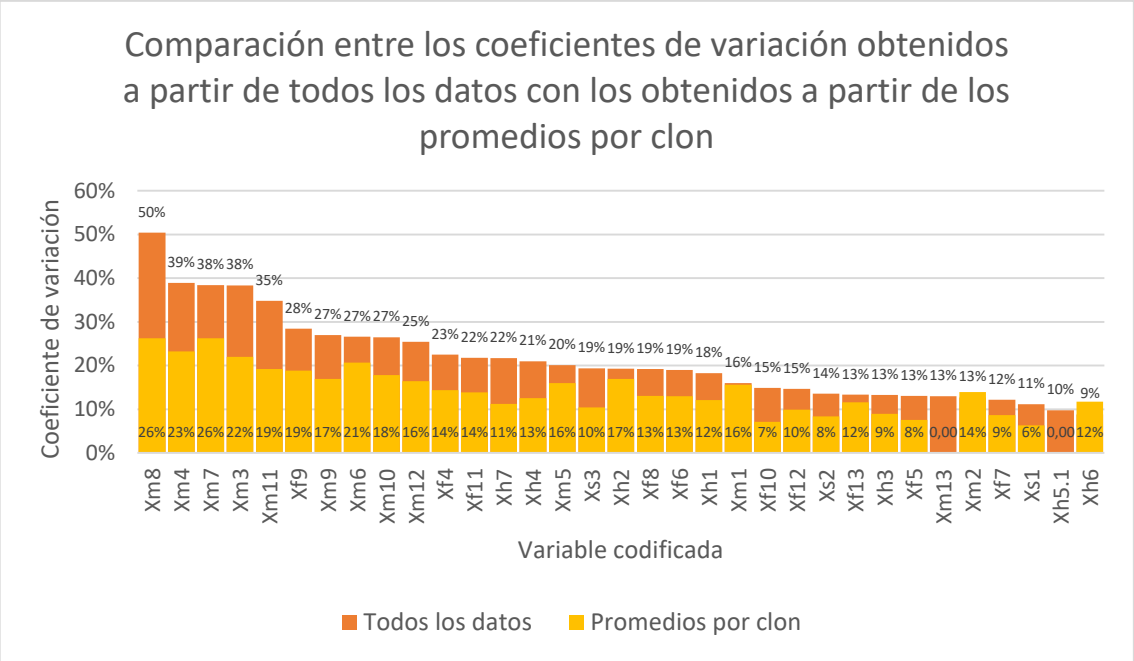

Gráfico 2-3: Comparación entre los coeficientes de variación por variable (En el anterior estudio no se incluyeron Xm13 ni Xh5.1 de forma cuantitativa).

Realizado por: Gabriela J. Obregón O. 2018.

La más alta variabilidad corresponde a Xm8 con un valor de 50%, que se considera como una variabilidad alta, la cual podría deberse en parte a que el tiempo tardado en pesar las semillas desde que se cosecha la mazorca influye en esta variable, pues las semillas pierden humedad con el pasar del tiempo, lo cual disminuye su peso (no es posible pesar las semillas de diferentes mazorcas simultáneamente). Por otra parte, se debe a que el número de semillas por mazorca sí es muy variable, pues el mínimo de semillas encontradas fue 5 y el máximo 68. Le siguen las variables Xm4, Xm7, Xm3 y Xm11, con valores entre 35% y 39%. De acuerdo a la codificación de color de la **Tabla 1-3** se aprecia que de manera general la unidad de análisis más variable es la mazorca, mientras que hoja, flor y semilla obtuvieron valores más bajos, es decir, menor variabilidad.

Las variables que presentaron menor variabilidad fueron Xh6, Xh5.1, Xs1, Xf7, Xm2, Xm13, Xf5, Xh3, Xf13, y Xs2, con coeficientes de variación por debajo del 15%.

En el **Gráfico 3-3** se muestra una comparación entre los coeficientes de variación obtenidos de todos los datos y los obtenidos únicamente a partir de los promedios de cada clon, donde se aprecia que, al ordenarlos de mayor a menor, si bien la tendencia de orden descendente es similar, los primeros en general presentan valores mayores que los segundos, lo cual es un indicio de que conviene utilizar todos los datos para reflejar la variabilidad total de forma más realista.

3.2 Pruebas no paramétricas.

Tabla 2-3: Valores-P resultantes de las pruebas de Kruskal-Wallis y ANOVA basado en rango para cada variable.

Variable	Valor-P	
	Kruskal-Wallis	**ANOVA basado en rango**
1. Xm1	0.146	0.000
2. Xm2	0.000	0.000
3. Xm3	0.198	0.000
4. Xm4	0.185	0.000
5. Xm5	0.000	0.000
6. Xm6	0.000	0.000
7. Xm7	0.000	0.000
8. Xm8	0.105	0.000
9. Xm9	0.036	0.000
10. Xm10	0.131	0.000
11. Xm11	0.012	0.000
12. Xm12	0.000	0.000
13. Xm13	0.144	0.000
14. Xs1	0.000	0.000
15. Xs2	0.000	0.000
16. Xs3	0.000	0.000
17. Xh1	0.000	0.000
18. Xh2	0.027	0.000
19. Xh3	0.162	0.000
20. Xh4	0.000	0.000
21. Xh5.1	0.465	0.000
22. Xh6	0.000	0.000
23. Xh7	0.001	0.000
24. Xf4	0.000	0.000
25. Xf5	0.000	0.000
26. Xf6	0.000	0.000
27. Xf7	0.000	0.000

28. Xf8	0.000	0.000
29. Xf9	0.000	0.000
30. Xf10	0.000	No disponible
31. Xf11	0.005	0.000
32. Xf12	0.000	0.000
33. Xf13	0.091	0.000

Realizado por: Gabriela J. Obregón O. 2018

En la **Tabla 2-3** se presentan los valores-p que permiten determinar las variables para las cuales se concluyó que existe suficiente evidencia estadística para rechazar la igualdad de todos los clones, y concluir que al menos un par de clones difiere significativamente, con una y otra prueba.

Mientras que Kruskal-Wallis detectó que sólo 25 variables son significativas, el ANOVA basado en rangos detectó que todas las variables lo son. Con esto se corrobora que las variables en estudio son adecuadas para discriminar los clones de forma univariada. Se adjunta en la *nube* (link: https://onedrive.live.com/?authkey=%21AIJE2OcZQnWrOgE&id=CA77A5953C07ED0C%21140440&cid=CA77A5953C07ED0C) las comparaciones entre pares de clones para cada variable con el método de Tukey (si el intervalo contiene al cero la diferencia es no significativa, pero si no lo contiene, se detectó como significativa).

3.3 Variables redundantes

En la matriz de Mazorcas los autovalores de la matriz de covarianzas son:

λ_{M1}	λ_{M2}	λ_{M3}	λ_{M4}	λ_{M5}	λ_{M6}	λ_{M7}	λ_{M8}	λ_{M9}	λ_{M10}	λ_{M11}	λ_{M12}	λ_{M13}
92064.43	2337.27	168.9972	75.97	11.25	3.38	2.96	0.40	0.25	0.04	0.03	0.000911	0.00000076

Se observan 2 autovalores próximos a 0, cuyos autovectores asociados proporcionan los coeficientes de las combinaciones lineales. De esta forma se obtuvo que:

La variable X_{m9} se puede expresar como: $X_{m9} = -W_0 + 0.9999562X_{m10} + 0.9999997X_{m11}$

Lo cual se explica debido a que la forma de calcular la variable X_{m11} (Peso de pulpa y testa de 5 semillas) fue como una diferencia entre las variables X_{m9} y X_{m10} (Peso de 5 semillas húmedas con pulpa y testa y Peso de 5 semillas húmedas sin pulpa y testa, respectivamente).

La variable X_{m13} se puede expresar como: $X_{m13} = -W_0 + 0.111757X_{m1} - 0.230245X_{m2} + 0.013142X_{m5} + 0.0030668X_{m6}$

En la matriz de Semillas ningún autovalor fue aproximadamente nulo en comparación con los demás:

λ_{S1}	λ_{S2}	λ_{S3}
0.08829031	0.02856166	0.01713326

En la matriz de Hojas los autovalores son:

λ_{H1}	λ_{H2}	λ_{H3}	λ_{H4}	λ_{H5}	λ_{H6}	λ_{H7}
232.7628	146.4393	35.7244	1.3865	1.0008	0.006673	0.002356

Se observan 2 autovalores próximos a 0 en comparación con los demás autovalores. Con los coeficientes de los autovectores asociados se determinó que:

$$X_{h5} = -W_0 + 0.048664X_{h1} + 0.054148X_{h2} + 0.247876X_{h3} - 0.140925X_{h4}$$

$$X_{h3} = -W_0 + 0.110179X_{h1} - 0.249577X_{h2} - 0.04459X_{h4} - 0.249744X_{h5}$$

Así, además de las filas con valores extremos, las variables X_{m9}, X_{m13}, X_{h5} y X_{h3} se retiraron de sus respectivas matrices antes de comenzar con el análisis. Posteriormente se retiró también la variable X_{m3}, pues se descubrió que para varios clones esta variable fue calculada como una combinación lineal: $X_{m3} = X_{m4} + X_{m8}$, lo cual dificultaba el cálculo de las distancias de Mahalanobis entre poblaciones al evitar que la matriz de covarianzas por clon sea invertible.

Las nuevas dimensiones de las matrices son:

Mazorcas: 599×10

Semillas: 2998×3

Hojas: 894×5

3.4 AMOVA

Tabla 3-3: Análisis de Varianza Molecular.

Fuente de variación	Grados de libertad	Suma de cuadrados	Porcentaje de variación
Entre poblaciones	2	15.285	1.6%
Entre individuos dentro de las poblaciones	27	185.615	3.5%
Dentro de individuos	30	192.000	94.9%
Total	59	392.900	100%

Realizado por: Gabriela J. Obregón O. 2018

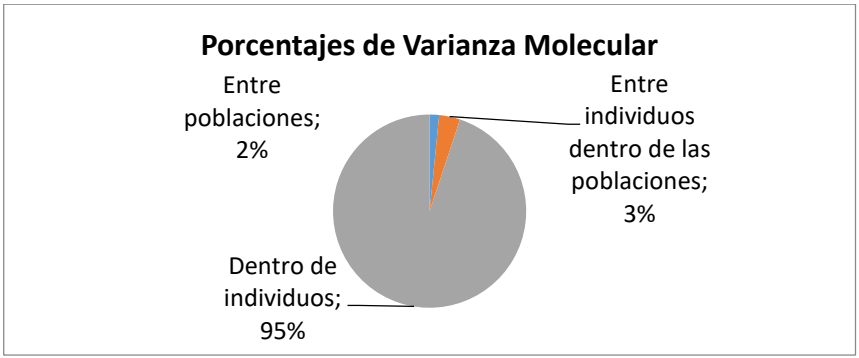

Gráfico 3-3: Porcentaje de variación a nivel molecular.

Realizado por: Gabriela J. Obregón O. 2018.

Los resultados de la **Tabla 3-3**, que se muestran en el **Gráfico 3-3**, indica que el mayor porcentaje de variación se atribuye a diferencias propias entre individuos (94.9%), seguido de la variación entre las poblaciones (1.6%). La variación entre individuos dentro de poblaciones de 3.5% es baja.

3.5 Análisis descriptivo multivariado

3.5.1 Matriz de mazorcas

Tabla 4-3: Matriz de covarianzas de las variables de mazorca.

	Xm1	*Xm2*	*Xm4*	*Xm5*	*Xm6*	*Xm7*	*Xm8*	*Xm10*	*Xm11*	*Xm12*
Xm1	8.63	2.03	411.17	0.32	0.09	20.22	113.97	1.71	3.35	1.36
Xm2	2.03	1.28	190.21	0.18	0.09	5.43	41.66	1.22	1.56	0.83
Xm4	411.17	190.21	36916.46	32.86	18.43	933.69	6995.47	190.45	256.51	127.63
Xm5	0.32	0.18	32.86	0.07	0.03	-0.13	2.96	0.24	0.25	0.16
Xm6	0.09	0.09	18.43	0.03	0.05	-0.22	0.43	0.11	0.08	0.08
Xm7	20.22	5.43	933.69	-0.13	-0.22	146.13	512.46	-7.00	-0.24	-2.47
Xm8	113.97	41.66	6995.47	2.96	0.43	512.46	3191.62	37.20	71.25	28.44
Xm10	1.71	1.22	190.45	0.24	0.11	-7.00	37.20	6.47	4.24	3.58
Xm11	3.35	1.56	256.51	0.25	0.08	-0.24	71.25	4.24	8.23	2.59
Xm12	1.36	0.83	127.63	0.16	0.08	-2.47	28.44	3.58	2.59	2.48

Realizado por: Gabriela J. Obregón O. 2018.

Tabla 5-3: Matriz de correlación de las variables de mazorca.

	Xm1	*Xm2*	*Xm4*	*Xm5*	*Xm6*	*Xm7*	*Xm8*	*Xm10*	*Xm11*	*Xm12*
Xm1	1.00	0.61	0.73	0.40	0.14	0.57	0.69	0.23	0.40	0.29
Xm2	0.61	1.00	0.88	0.57	0.38	0.40	0.65	0.43	0.48	0.47
Xm4	0.73	0.88	1.00	0.63	0.43	0.40	0.64	0.39	0.47	0.42
Xm5	0.40	0.57	0.63	1.00	0.53	-0.04	0.19	0.35	0.32	0.37
Xm6	0.14	0.38	0.43	0.53	1.00	-0.08	0.03	0.19	0.13	0.22
Xm7	0.57	0.40	0.40	-0.04	-0.08	1.00	0.75	-0.23	-0.01	-0.13
Xm8	0.69	0.65	0.64	0.19	0.03	0.75	1.00	0.26	0.44	0.32
Xm10	0.23	0.43	0.39	0.35	0.19	-0.23	0.26	1.00	0.58	0.89
Xm11	0.40	0.48	0.47	0.32	0.13	-0.01	0.44	0.58	1.00	0.57
Xm12	0.29	0.47	0.42	0.37	0.22	-0.13	0.32	0.89	0.57	1.00

Relación lineal	Alta	Moderada	Baja	Débil
Rango	(+-) 0.75-1.00	(+-) 0.50-0.75	(+-) 0.25-0.50	(+-) 0.00-0.25

Realizado por: Gabriela J. Obregón O. 2018.

Varianza generalizada: $|\mathbf{S}| = 22339054.51$ El grado de dispersión en el espacio es grande.

Varianza total: $tr(\mathbf{S}) = 40281.42$

Coeficiente de dependencia: $\eta^2 = 0.99976$

El valor próximo a 1 indica que hay relaciones lineales entre las variables.

Correlación lineal alta:

A mayor peso de 5 semillas húmedas sin pulpa y testa, mayor peso de 5 semillas secas (0.89). A mayor ancho de mazorca, mayor peso de cáscara (0.88). A mayor número total de semillas, mayor peso total de semillas (0.75).

Correlación lineal moderada:

A mayor largo de mazorca, mayor peso de cáscara (0.73). A mayor largo de mazorca, mayor peso total de semillas (0.69). A mayor ancho de mazorca, mayor peso total de semillas (0.65). A mayor peso de cáscara, mayor peso total de semillas (0.64). A mayor peso de cáscara, mayor espesor de cáscara en el lomo (0.63). A mayor largo de mazorca, mayor ancho de mazorca (0.61). A mayor peso de 5 semillas húmedas sin pulpa y testa, mayor peso de pulpa y testa de 5 semillas (0.58). A mayor peso de pulpa y testa de 5 semillas, mayor peso de 5 semillas secas (0.57). A mayor ancho de mazorca, mayor espesor de cáscara en el lomo (0.57). A mayor largo de mazorca, mayor número total de semillas (0.57).

Correlaciones lineales negativas (todas débiles):

A mayor número total de semillas, menor peso de 5 semillas húmedas sin pulpa y testa (-0.23). A mayor número total de semillas, menor peso de 5 semillas secas (-0.13). A mayor espesor de cáscara en el surco, menor número total de semillas (-0.08). A mayor espesor de cáscara en el lomo, menor número total de semillas (-0.04). A mayor número total de semillas, menor peso de pulpa y testa de 5 semillas (-0.01).

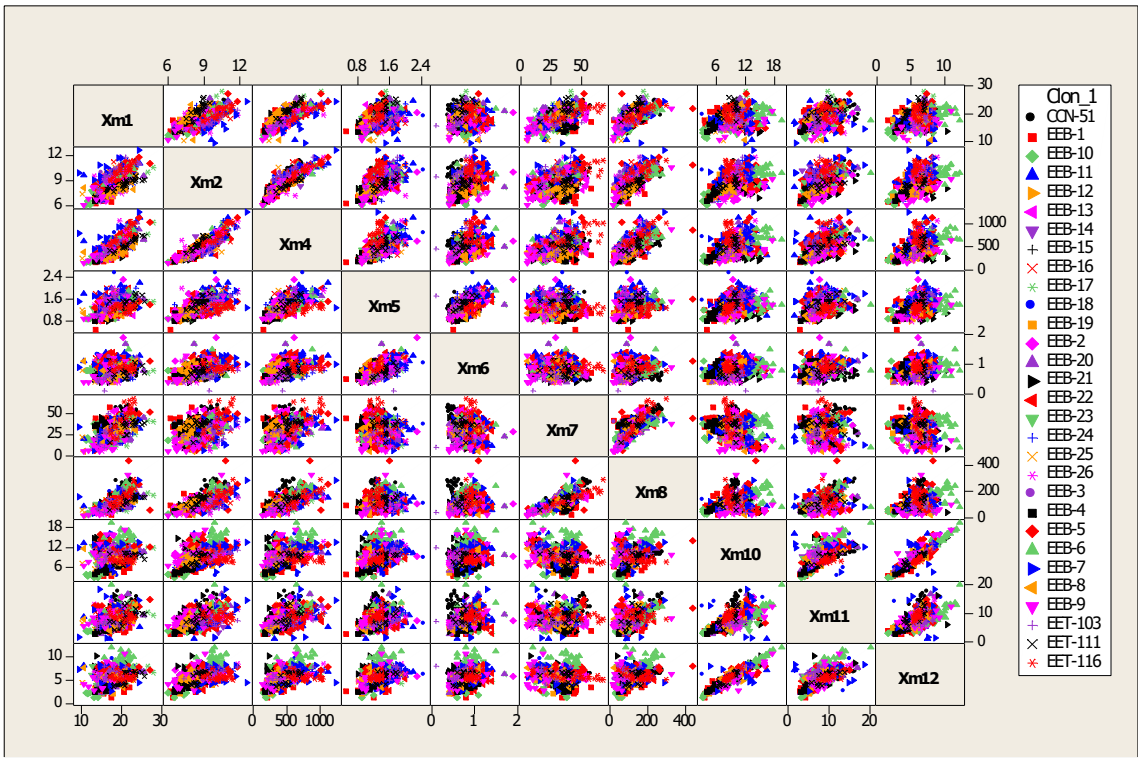

Gráfico 4-3: Gráficos de dispersión para visualizar correlaciones entre pares de variables de mazorca.

Realizado por: Gabriela J. Obregón O. 2018.

3.5.2 Matriz de semillas

Tabla 6-3: Matriz de covarianzas de variables de semilla.

	Xs1	*Xs2*	*Xs3*
Xs1	0.07	0.03	0.01
Xs2	0.03	0.03	0.01
Xs3	0.01	0.01	0.03

Realizado por: Gabriela J. Obregón O. 2018.

Varianza generalizada: $|\mathbf{S}| = 0.000043$ El grado de dispersión en el espacio es pequeño.

Varianza total: $tr(\mathbf{S}) = 0.13398$

Matriz de correlación

Tabla 7-3: Matriz de correlación entre variables de semilla.

	Xs1	*Xs2*	*Xs3*
Xs1	1.00	0.54	0.29
Xs2	0.54	1.00	0.38
Xs3	0.29	0.38	1.00

Relación lineal	Alta	Moderada	Baja	Débil
Rango	(+-) 0.75-1.00	(+-) 0.50-0.75	(+-) 0.25-0.50	(+-) 0.00-0.25

Realizado por: Gabriela J. Obregón O. 2018.

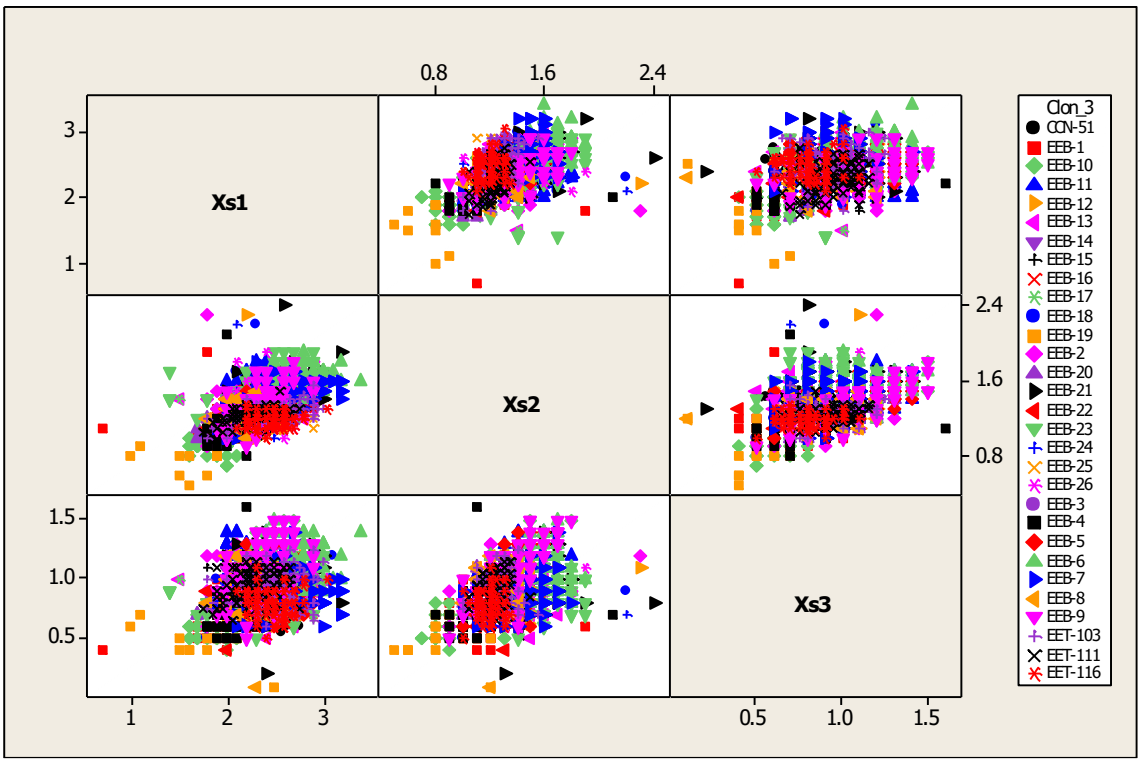

Gráfico 5-3: Gráficos de dispersión para visualizar correlaciones entre pares de variables de semilla.

Realizado por: Gabriela J. Obregón O. 2018.

Coeficiente de dependencia: $\eta^2 = 0.3999$

El valor lejano a 1 indica que las variables no tienen relaciones lineales fuertes.

Correlación lineal moderada:

A mayor largo de semilla, mayor ancho de semilla (0.54).

3.5.3 Matriz de hojas

Tabla 8-3: Matriz de covarianzas de variables de hoja.

	Xh1	*Xh2*	*Xh4*	*Xh6*	*Xh7*
Xh1	30.8	8.9	15.3	14.0	-19.3
Xh2	8.9	4.5	4.8	12.8	5.9
Xh4	15.3	4.8	9.5	5.4	-5.6
Xh6	14.0	12.8	5.4	185.8	45.7
Xh7	-19.3	5.9	-5.6	45.7	186.5

Realizado por: Gabriela J. Obregón O. 2018.

Varianza generalizada: $|\mathbf{S}| = 1540349.76$ El grado de dispersión en el espacio es grande.

Varianza total: $tr(\mathbf{S}) = 417.1$

Tabla 9-3: Matriz de correlación entre variables de hoja.

	Xh1	*Xh2*	*Xh4*	*Xh6*	*Xh7*
Xh1	1.00	0.75	0.90	0.18	-0.26
Xh2	0.75	1.00	0.73	0.44	0.20
Xh4	0.90	0.73	1.00	0.13	-0.13
Xh6	0.18	0.44	0.13	1.00	0.25
Xh7	-0.26	0.20	-0.13	0.25	1.00

Relación lineal	Alta	Moderada	Baja	Débil
Rango	(+-) 0.75-1.00	(+-) 0.50-0.75	(+-) 0.25-0.50	(+-) 0.00-0.25

Realizado por: Gabriela J. Obregón O. 2018.

Coeficiente de dependencia: $\eta^2 = 0.9666$

El valor próximo a 1 indica que hay relaciones lineales entre las variables.

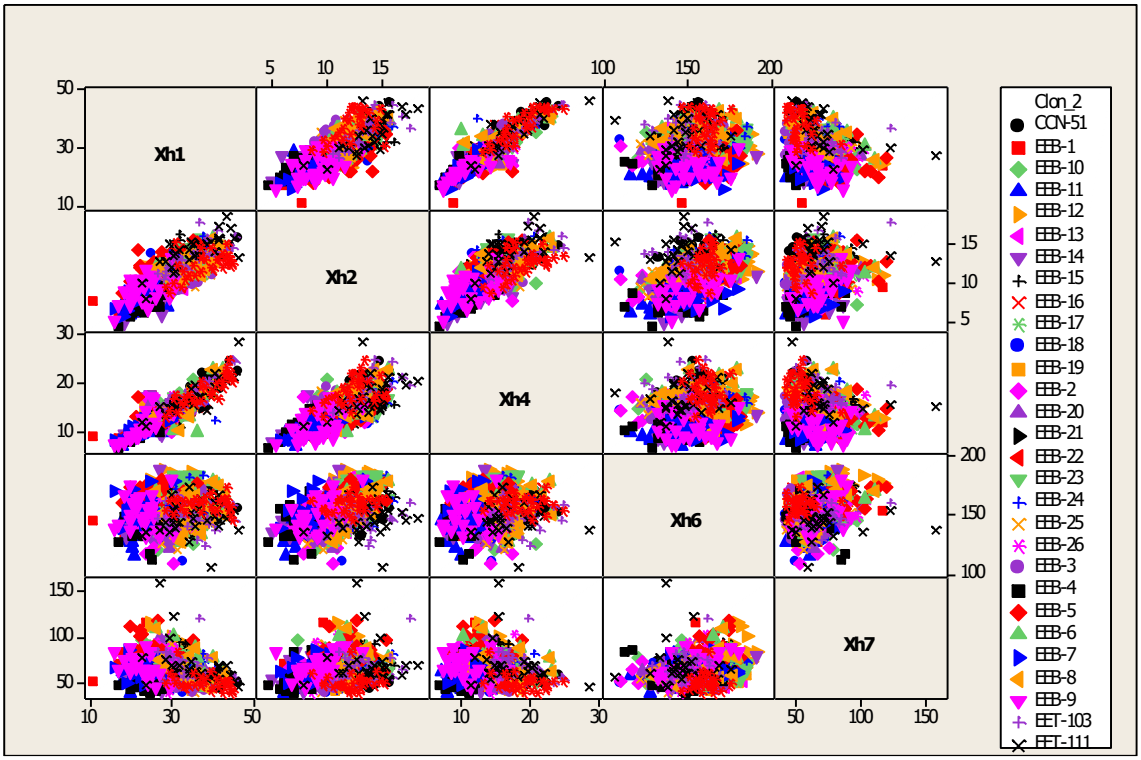

Gráfico 6-3: Gráficos de dispersión entre pares de variables de hoja.

Realizado por: Gabriela J. Obregón O. 2018.

Correlación lineal alta:

A mayor largo de hoja, mayor largo desde la base hasta el punto más ancho del limbo (0.90). A mayor largo de hoja, mayor ancho de hoja (0.75).

Correlación lineal moderada:

A mayor ancho de hoja, mayor largo desde la base hasta el punto más ancho del limbo (0.73).

Correlación lineal negativa:

A mayor largo de hoja, menor ángulo apical (-0.26 – baja). A mayor largo desde la base hasta el punto más ancho del limbo, menor ángulo apical (-0.13 – débil).

3.5.4 Matriz de flores

Varianza generalizada: $|\mathbf{S}| = 0.000011$ El grado de dispersión en el espacio es pequeño.

Varianza total: $tr(\mathbf{S}) = 6.25$

Coeficiente de dependencia: $\eta^2 = 0.9928$

El valor próximo a 1 indica que hay relaciones lineales entre las variables.

Tabla 10-3: Matriz de covarianzas de variables de flor.

	Xf1.1	*Xf2.1*	*Xf3.1*	*Xf4.1*	*Xf5.1*	*Xf6.1*	*Xf7.1*	*Xf8.1*	*Xf9.1*	*Xf10.1*
Xf1.1	0.55	0.12	0.27	0.10	0.13	0.30	0.02	0.37	0.42	0.09
Xf2.1	0.12	0.37	0.07	0.18	-0.05	0.05	-0.06	0.27	-0.05	0.05
Xf3.1	0.27	0.07	0.43	0.04	0.24	0.25	0.11	0.23	0.57	-0.11
Xf4.1	0.10	0.18	0.04	0.55	-0.03	0.06	-0.07	0.16	0.23	0.08
Xf5.1	0.13	-0.05	0.24	-0.03	0.43	0.26	0.08	0.15	0.31	-0.09
Xf6.1	0.30	0.05	0.25	0.06	0.26	0.43	0.16	0.31	0.46	-0.08
Xf7.1	0.02	-0.06	0.11	-0.07	0.08	0.16	0.24	0.07	0.28	-0.12
Xf8.1	0.37	0.27	0.23	0.16	0.15	0.31	0.07	0.61	0.30	0.04
Xf9.1	0.42	-0.05	0.57	0.23	0.31	0.46	0.28	0.30	2.05	0.02
Xf10.1	0.09	0.05	-0.11	0.08	-0.09	-0.08	-0.12	0.04	0.02	0.58

Realizado por: Gabriela J. Obregón O. 2018.

Tabla 11-3: Matriz de correlación entre variables de flor.

	Xf1.1	*Xf2.1*	*Xf3.1*	*Xf4.1*	*Xf5.1*	*Xf6.1*	*Xf7.1*	*Xf8.1*	*Xf9.1*	*Xf10.1*
Xf1.1	1.00	0.27	0.55	0.18	0.28	0.62	0.04	0.64	0.39	0.16
Xf2.1	0.27	1.00	0.18	0.39	-0.13	0.12	-0.19	0.58	-0.06	0.10
Xf3.1	0.55	0.18	1.00	0.09	0.54	0.58	0.36	0.44	0.60	-0.22
Xf4.1	0.18	0.39	0.09	1.00	-0.05	0.13	-0.21	0.29	0.22	0.15
Xf5.1	0.28	-0.13	0.54	-0.05	1.00	0.60	0.24	0.30	0.32	-0.19
Xf6.1	0.62	0.12	0.58	0.13	0.60	1.00	0.49	0.61	0.48	-0.16
Xf7.1	0.04	-0.19	0.36	-0.21	0.24	0.49	1.00	0.18	0.40	-0.32
Xf8.1	0.64	0.58	0.44	0.29	0.30	0.61	0.18	1.00	0.27	0.07
Xf9.1	0.39	-0.06	0.60	0.22	0.32	0.48	0.40	0.27	1.00	0.02
Xf10.1	0.16	0.10	-0.22	0.15	-0.19	-0.16	-0.32	0.07	0.02	1.00

Relación lineal	Alta	Moderada	Baja	Débil
Rango	(+-) 0.75-1.00	(+-) 0.50-0.75	(+-) 0.25-0.50	(+-) 0.00-0.25

Realizado por: Gabriela J. Obregón O. 2018.

Correlación lineal moderada:

A mayor promedio de ancho de sépalo en 20 flores observadas, mayor promedio de largo de pedúnculo en 10 flores observadas (0.64). A mayor promedio de ancho de sépalo en 20 flores

observadas, mayor promedio de largo del ovario en 10 flores observadas (0.62). A mayor promedio de largo del ovario en 10 flores observadas, promedio de largo de pedúnculo en 10 flores observadas (0.61). A mayor promedio de ancho de lígula en 15 flores observadas, mayor promedio de largo del estaminoide en 6 flores observadas (0.60). Promedio de largo del estilo en 10 flores observadas, mayor promedio de largo del ovario en 10 flores observadas (0.60). A mayor promedio de ancho de lígula en 15 flores observadas, mayor promedio de largo del ovario en 10 flores observadas (0.58). A mayor, promedio de ancho de sépalo en 20 flores observadas, mayor promedio de ancho de lígula en 15 flores observadas (0.55). A mayor promedio de ancho de lígula en 15 flores observadas, mayor promedio de largo del estilo en 10 flores observadas (0.54).

Correlación lineal negativa baja:

A mayor promedio de óvulos observados en el ovario de 5 flores, menor promedio de ancho del ovario en 10 flores observadas.

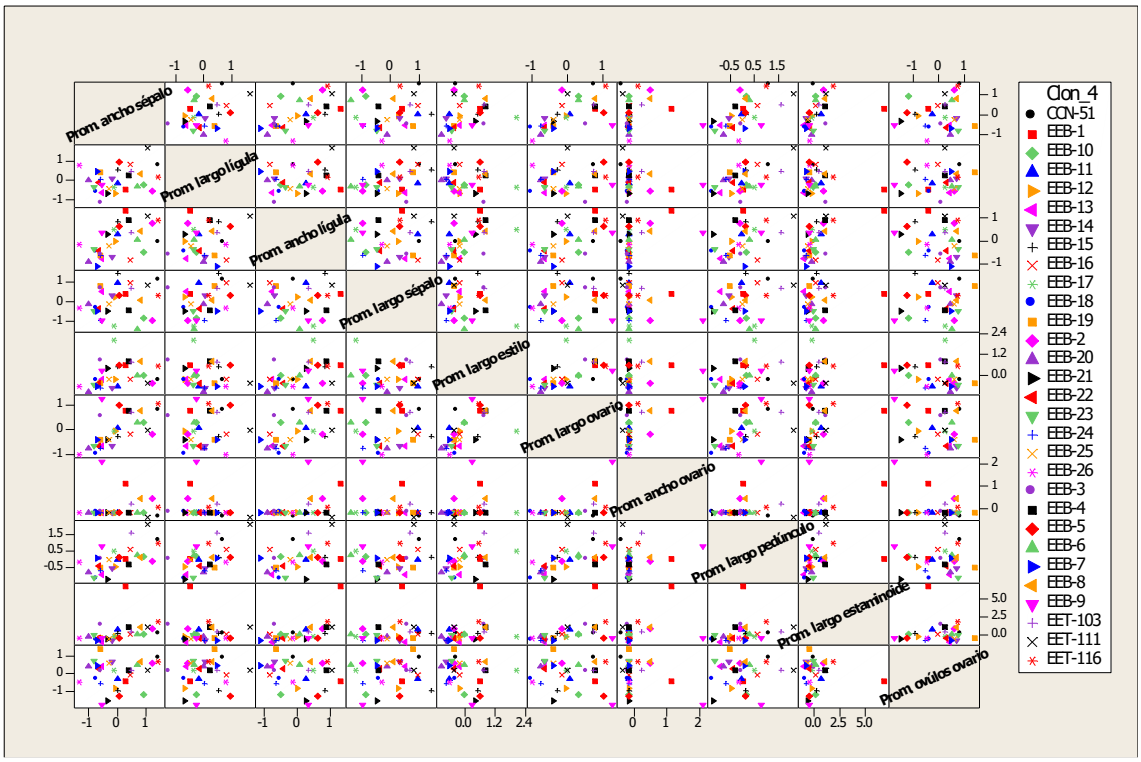

Gráfico 7-3: Gráficos de dispersión para visualizar la correlación entre pares de variables.

Realizado por: Gabriela J. Obregón O. 2018.

3.6 Análisis de Coordenadas Principales

En los siguientes puntos, se muestran los porcentajes de variabilidad explicados por cada coordenada principal para cada una de las configuraciones, así como las gráficas correspondientes

a las primeras coordenadas. En los gráficos bidimensionales las etiquetas resaltadas en color amarillo corresponden a los testigos, y las etiquetas resaltadas en color verde corresponden a los 6 clones tipo Criollo que destacan en rendimiento.

3.6.1 Configuración Molecular

Las 3 primeras coordenadas principales explican el 32% de la variabilidad total.

En el **Gráfico 8-3** se aprecia cómo según los marcadores moleculares, casi todos los materiales de tipo Criollo se agrupan hacia la izquierda y los clones testigo están hacia la derecha. Los clones testigo EET-116 y CCN-51 se diferencian claramente del resto, y el único clon que se encuentra distante con el resto de Criollos es C23 que se asemeja más al forastero EET-116 y al híbrido CCN-51.

Tabla 12-3: Variabilidad explicada por las coordenadas principales de la configuración molecular.

Coordenada Principal - Dimensión q (q=1, …, 30)	Autovalor	Variabilidad geométrica explicada por la coordenada q	Fracción de variabilidad explicada por la coordenada q	Porcentaje de variabilidad explicada hasta los q primeros ejes
1	1.8825	0.063	0.12	12%
2	1.6909	0.056	0.11	23%
3	1.4708	0.049	0.09	32%
4	1.1806	0.039	0.08	40%
5	0.9033	0.030	0.06	46%
6	0.8281	0.028	0.05	51%
7	0.8103	0.027	0.05	56%
8	0.7497	0.025	0.05	61%
9	0.6902	0.023	0.04	65%
10	0.6116	0.020	0.04	69%
11	0.5476	0.018	0.04	73%
12	0.4578	0.015	0.03	76%
13	0.4323	0.014	0.03	79%
14	0.3951	0.013	0.03	81%
15	0.3697	0.012	0.02	83%
16	0.3327	0.011	0.02	86%
17	0.3009	0.010	0.02	88%
18	0.2725	0.009	0.02	89%
19	0.2484	0.008	0.02	91%
20	0.2341	0.008	0.02	92%
21	0.2210	0.007	0.01	94%
22	0.2063	0.007	0.01	95%
23	0.1619	0.005	0.01	96%
24	0.1509	0.005	0.01	97%
25	0.1341	0.004	0.01	98%
26	0.1014	0.003	0.01	99%
27	0.0826	0.003	0.01	99%
28	0.0801	0.003	0.01	100%
29	0.0559	0.002	0.00	100%
30	0.0003	0.000	0.00	100%
Variabilidad geométrica		0.520	1.00	

Realizado por: Gabriela J. Obregón O. 2018.

Es notable además que CCN-51 se encuentra entre el forastero EET-116 y el trinitario EET-111, ya que es conocido que éstos fueron sus padres. En general los de tipo Criollo se asemejan más a los testigos EET-103 (siendo C19 el más similar con similitud 0.65) y EET-111 (siendo C8 y C19 los más similares con similitud 0.60). Las parejas de clones que presentaron menor similitud son: C16 y C23, C18 y EET-116, C21 y C23, con el mínimo coeficiente de similitud: 0.18. Los que presentaron mayor coeficiente de similitud fueron C16 y C21 con 0.90, seguidos de C4 y C10 con 0.88, C12 y C15 con 0.83, y C19 y C25 con 0.75.

En el **Gráfico 9-3** la agrupación es similar al gráfico anterior. En el **Gráfico 10-3** los clones de mayor similitud permanecen cercanos.

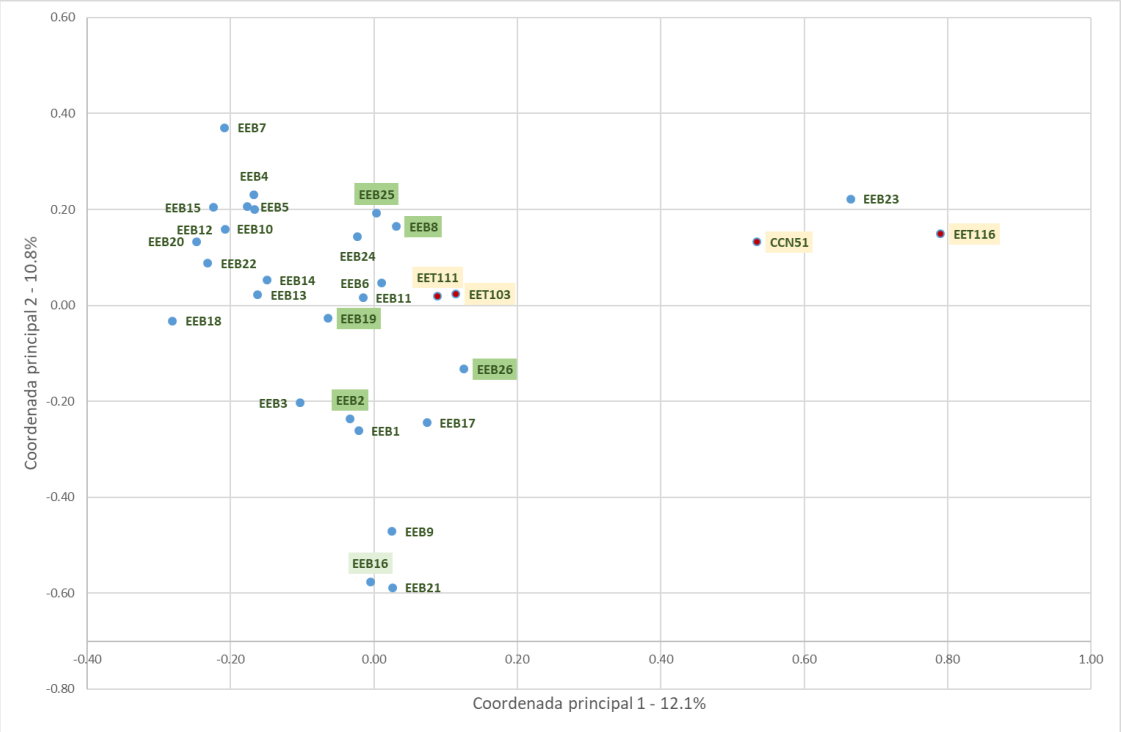

Gráfico 8-3: Coordenadas principales 1 y 2 de similitudes moleculares. Explican el 22.9% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

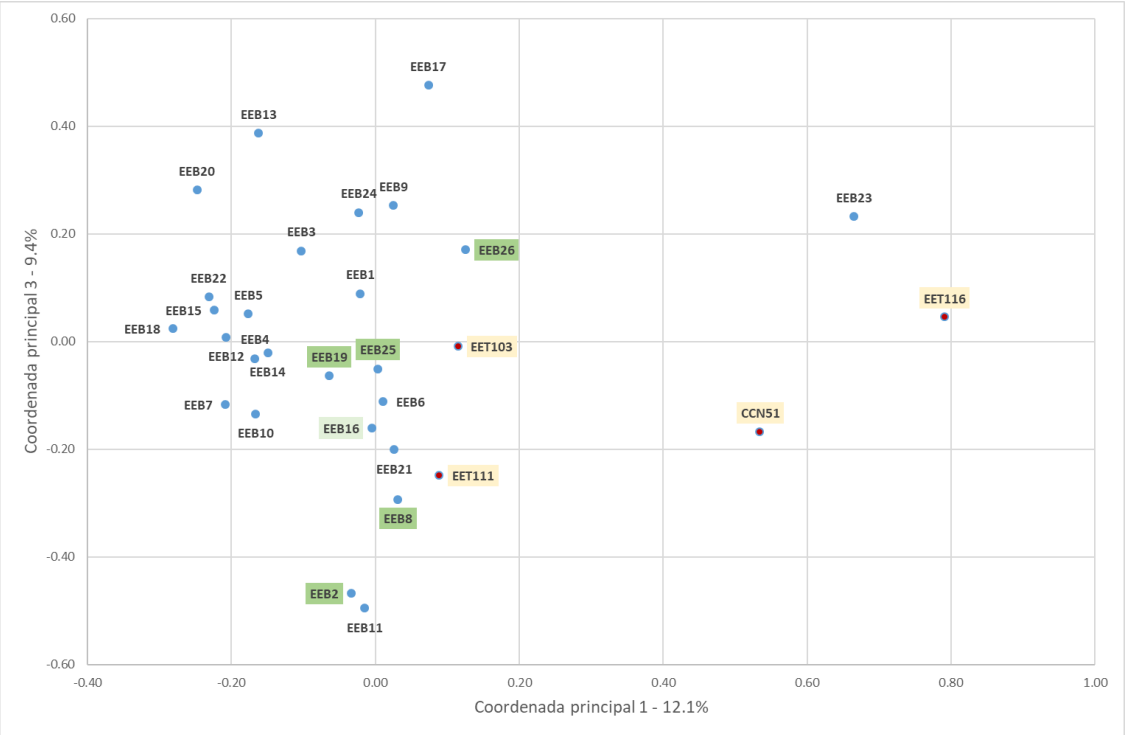

Gráfico 9-3: Coordenadas principales 1 y 3 de similitudes moleculares. Explican el 21.5% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

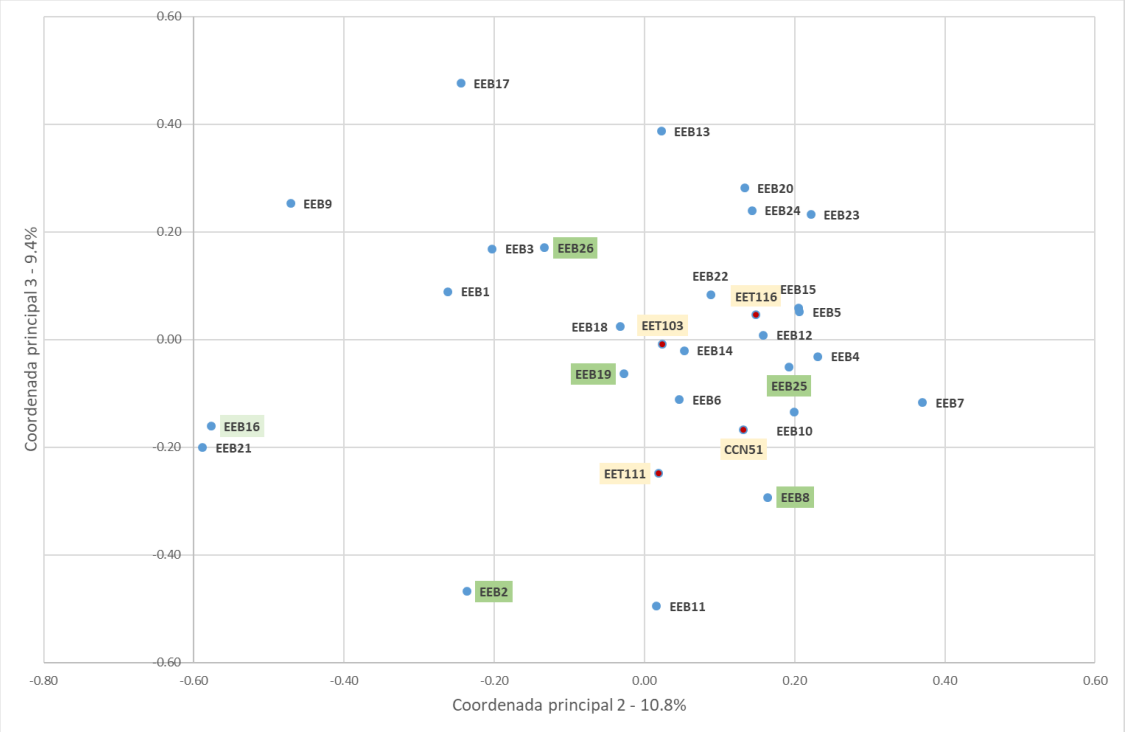

Gráfico 10-3: Coordenadas principales 2 y 3 de similitudes moleculares. Explican el 20.2% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

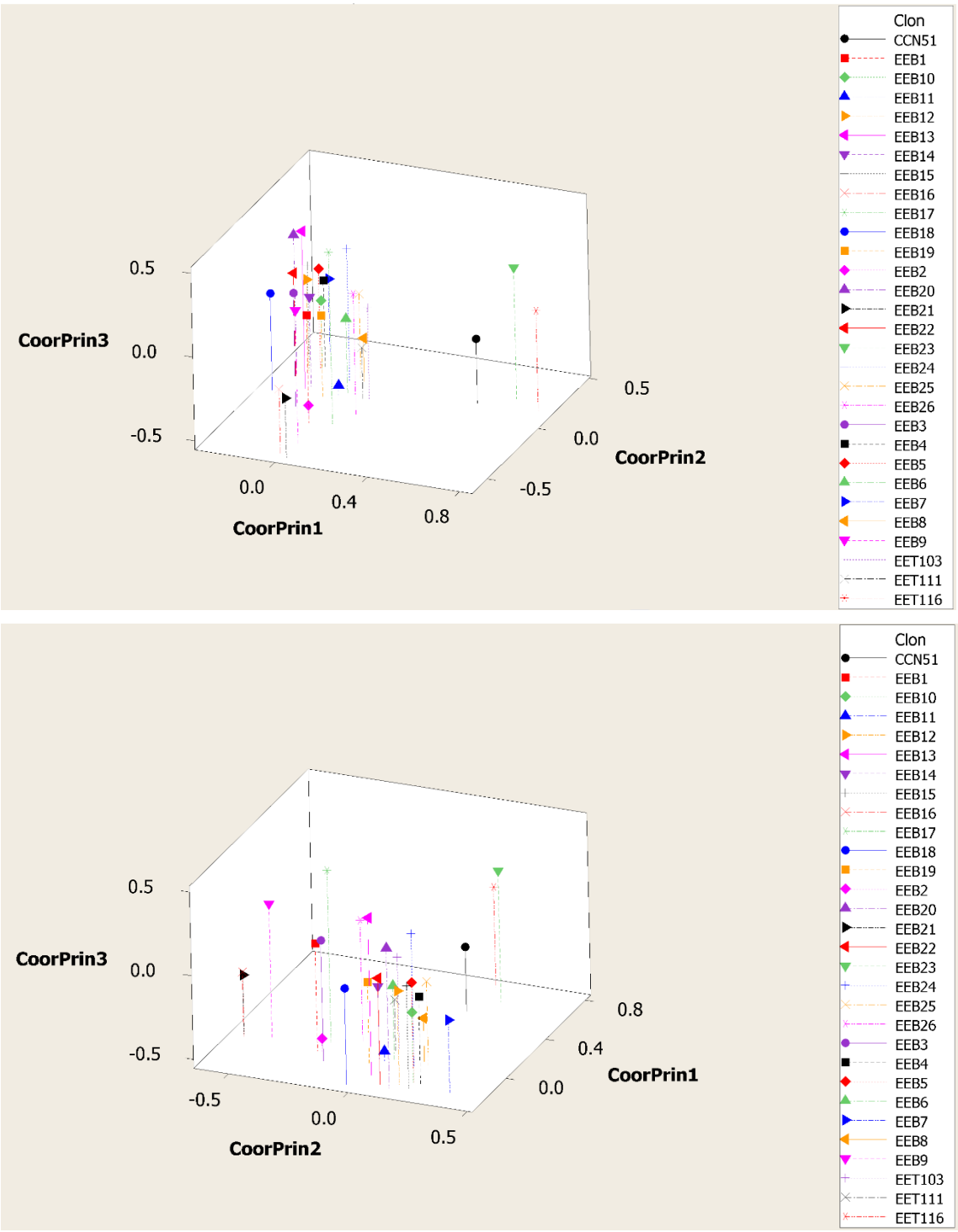

Gráfico 11-3: Visualización en 3D con las 3 primeras Coordenadas Principales de datos moleculares, desde diferentes perspectivas.

Realizado por: Gabriela J. Obregón O. 2018.

En la visualización en 3D (**Gráfico 11-3**) se aprecia la distribución en el espacio de los clones desde diferentes perspectivas. El porcentaje de variabilidad explicado es del 32%. Se aprecia

claramente cómo los de tipo Criollo conforman un grupo distanciado del Forastero EET-116, de C23 y de CCN-51. Además, se aprecia que C16, C21 y C9 se apartan un poco del resto.

Al enfocarse en los 6 materiales tipo Criollo de mayor rendimiento (**Gráfico 12-3**), se observa que C16 se distancia más de los demás, y que, considerando la similaridad, los clones más cercanos son C19 y C25 (0.75). Es notable además que los de mayor rendimiento (excepto C16) son molecularmente más similares al Nacional EET-103 y al Trinitario EET-111.

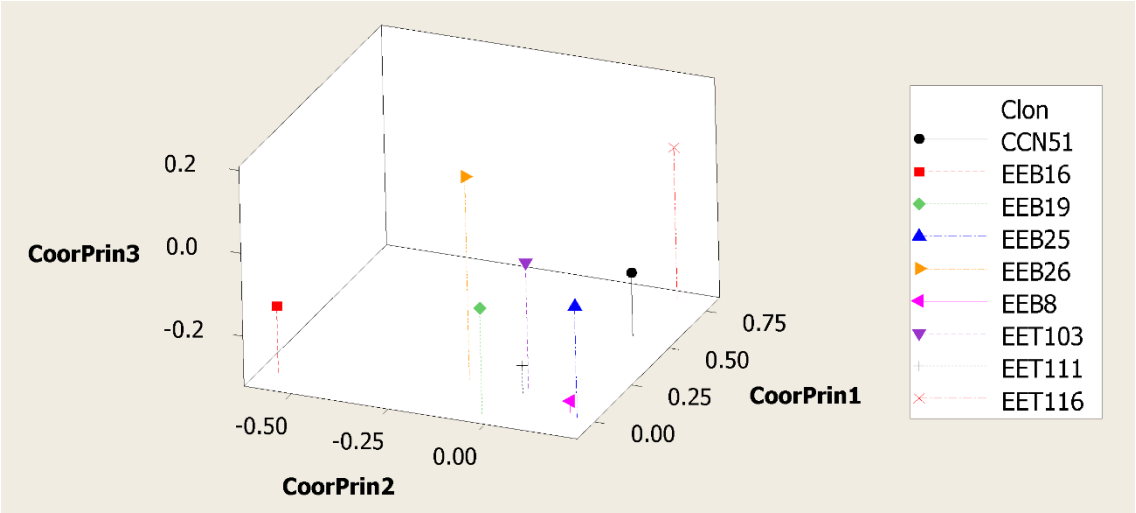

Gráfico 12-3: Gráfica en 3D con enfoque en los 6 de tipo Criollo de más alto rendimiento, en comparación con los testigos.

Realizado por: Gabriela J. Obregón O. 2018.

3.6.2 Configuración de Mazorca

Las 3 primeras coordenadas principales de los datos de mazorca explican el 23% de la variabilidad total.

Según los datos de mazorca se observa en los **Gráficos 13-3**, **14-3** y **15-3** cómo CCN-51 (que es altamente productivo) destaca manteniéndose distante de todos los demás. En el **Gráfico 13-3** destacan también los clones C5 y C6 por una parte, y C4 y C16 por otra, así como EET-111 y C1 en la parte superior. Los clones de alto rendimiento están cercanos, excepto C16. La mayoría de los clones tipo Criollo conforman un grupo diferenciado.

En el **Gráfico 14-3** conforman un grupo separado C7, C18, C22 y EET-116, otro grupo son C16, C4 y C14, y en la parte inferior C21 y C6. Los 5 de alto rendimiento permanecen cercanos.

Tabla 13-3: Variabilidad explicada por las coordenadas principales de la configuración de mazorca.

Coordenada Principal - Dimensión q (q=1, …, 29)	Autovalor	Variabilidad geométrica explicada por la coordenada q	Fracción de variabilidad explicada por la coordenada q	Porcentaje de variabilidad explicada hasta los q primeros ejes
1	228.3	7.6	0.10	10%
2	166.0	5.5	0.07	18%
3	123.8	4.1	0.06	23%
4	106.4	3.5	0.05	28%
5	99.6	3.3	0.04	32%
6	96.3	3.2	0.04	37%
7	87.6	2.9	0.04	40%
8	81.4	2.7	0.04	44%
9	79.3	2.6	0.04	48%
10	74.4	2.5	0.03	51%
11	72.9	2.4	0.03	54%
12	72.2	2.4	0.03	57%
13	71.3	2.4	0.03	61%
14	70.5	2.3	0.03	64%
15	69.5	2.3	0.03	67%
16	68.3	2.3	0.03	70%
17	67.2	2.2	0.03	73%
18	64.4	2.1	0.03	76%
19	63.6	2.1	0.03	79%
20	62.1	2.1	0.03	81%
21	60.7	2.0	0.03	84%
22	59.4	2.0	0.03	87%
23	57.7	1.9	0.03	89%
24	54.8	1.8	0.02	92%
25	53.5	1.8	0.02	94%
26	50.9	1.7	0.02	96%
27	42.0	1.4	0.02	98%
28	40.4	1.3	0.02	100%
29	0.0	0.0	0.00	100%
Variabilidad geométrica		74.8	1.00	

Realizado por: Gabriela J. Obregón O. 2018.

En el **Gráfico 15-3** se separan en un grupo C5 y C6, en otro C19, C4, C1 y EET-111 (Trinitario), y en la parte superior C18, C7, CCN-51, EET-116 (Forastero) y C22. Los 6 de alto rendimiento permanecen cercanos.

En 3D (**Gráfico 16-3**) se aprecia mejor cómo CCN-51 se aleja del resto. Un grupo podría ser C6 y C5. Otro grupo C1 y EET-111, y otro C4 y C16.

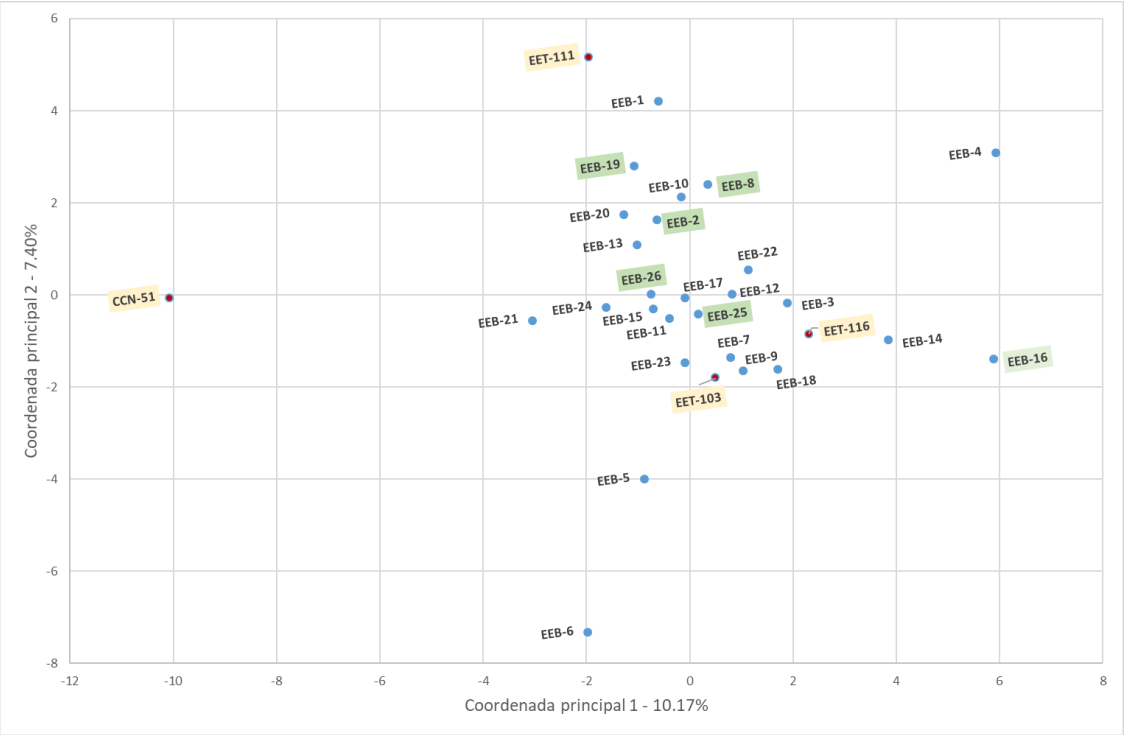

Gráfico 13-3: Coordenadas principales 1 y 2 de similitudes de mazorca. Explican el 17.6% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

Siguiendo a los clones de mayor rendimiento (**Gráfico 17-3**), están más cerca de los testigos EET-116 y EET-103. Se agrupan C2, C8 y C19, y C25 y C26. En general se mantienen cercanos. Se encuentran distantes de EET-111 (Trinitario).

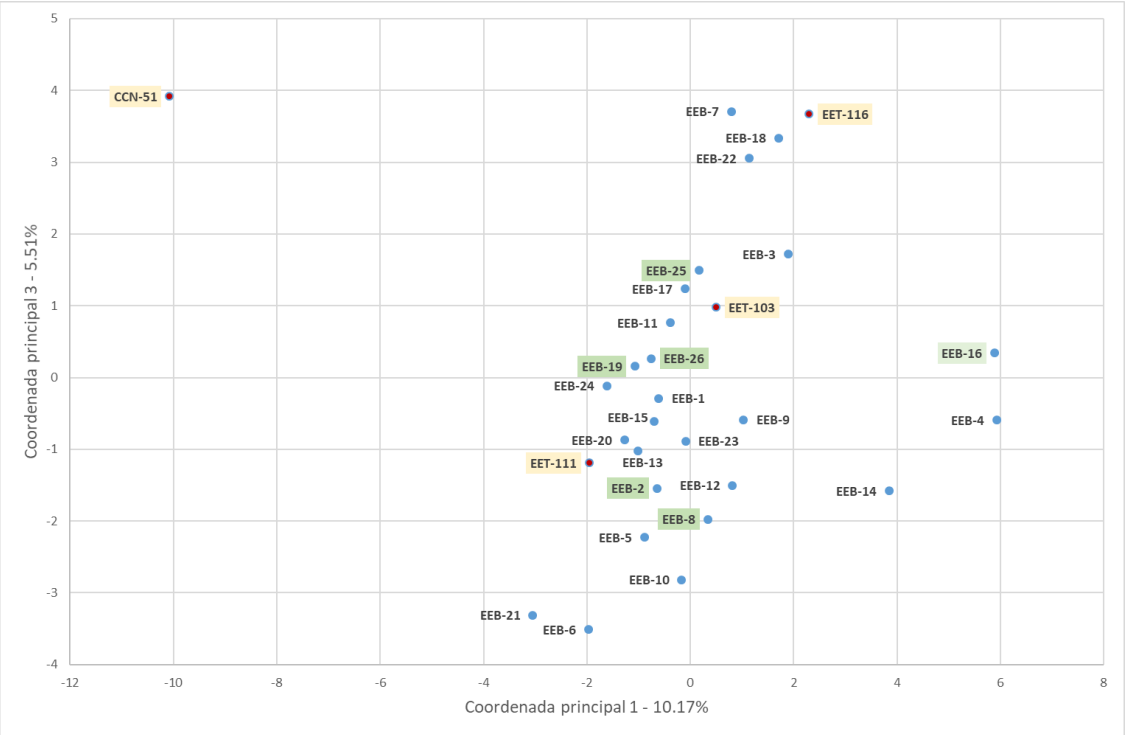

Gráfico 14-3: Coordenadas principales 1 y 3 de similitudes de mazorca. Explican el 15.7% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

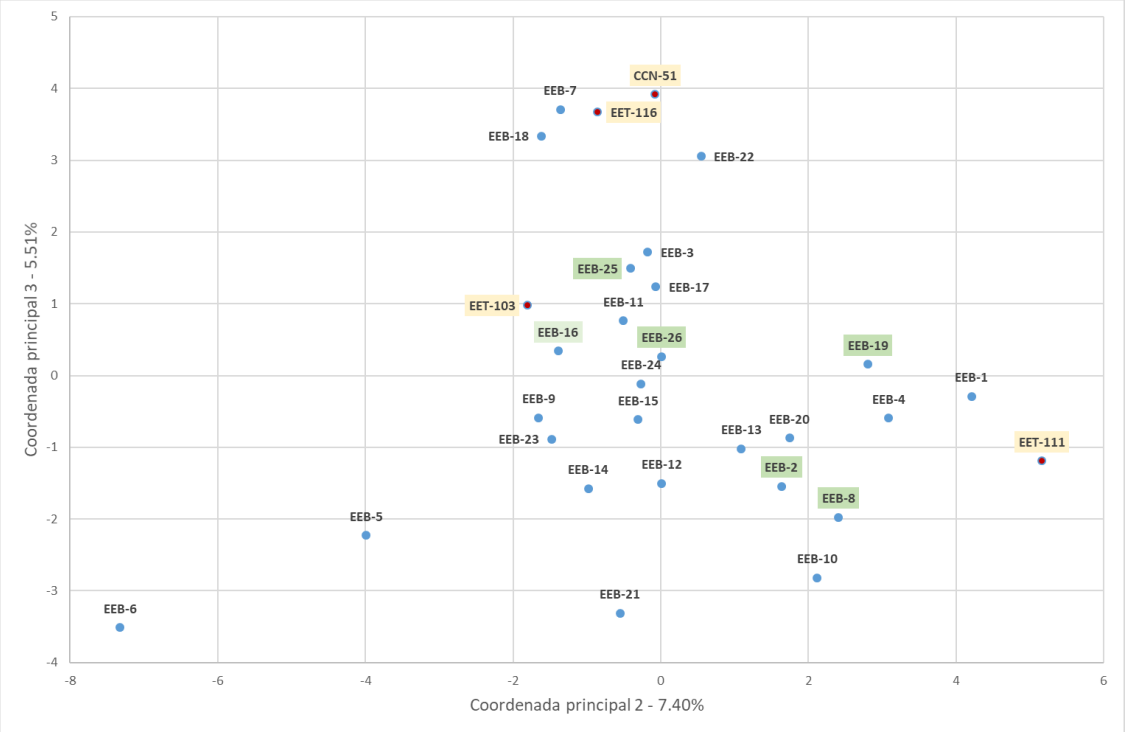

Gráfico 15-3: Coordenadas principales 2 y 3 de similitudes moleculares. Explican el 12.9% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

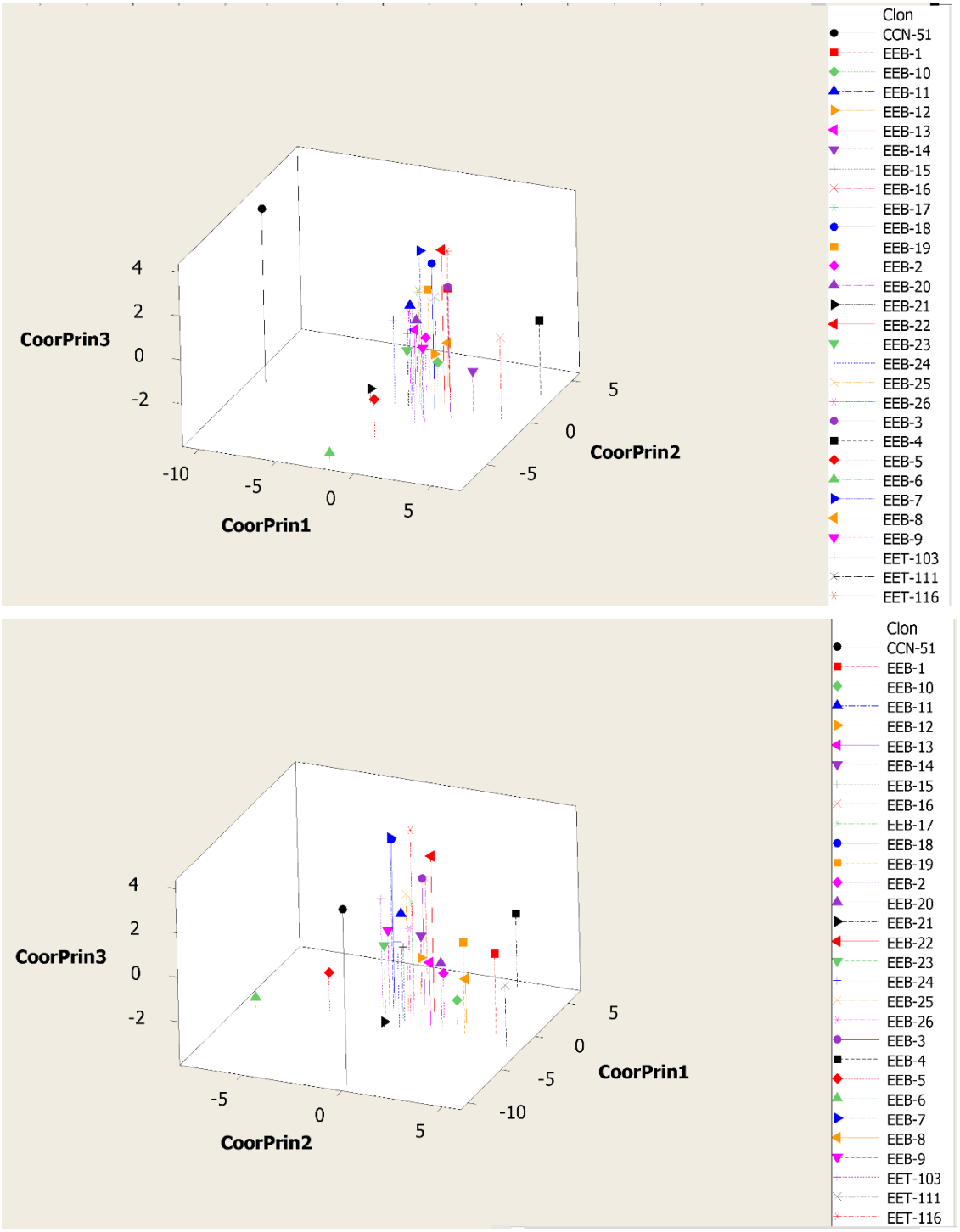

Gráfico 16-3: Visualización en 3D con las 3 primeras Coordenadas Principales de datos de mazorca, desde diferentes perspectivas.

Realizado por: Gabriela J. Obregón O. 2018.

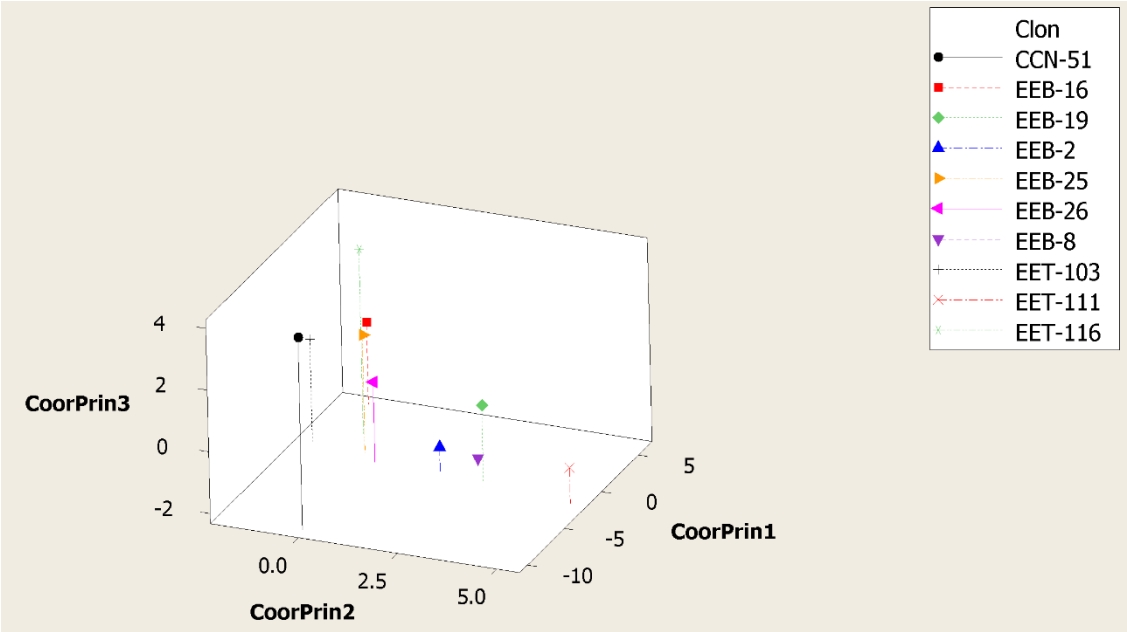

Gráfico 17-3: Gráfica en 3D con enfoque en los 6 de tipo Criollo de más alto rendimiento, en comparación con los testigos.

Realizado por: Gabriela J. Obregón O. 2018.

3.6.3 Configuración de Semilla

Las 3 primeras coordenadas principales explican el 27% de la variabilidad total.

En el **Gráfico 18-3** se observa que en semilla destacan los clones C3, C9 por un lado, y C6 por otro. CCN-51 se mantiene dentro del grupo de Criollos, sin destacar en semilla. Los testigos están dentro del mismo grupo de criollos, excepto EET-116. Otro grupo es C4, EET-116 y C1.

En el **Gráfico 19-3** y **20-3** destaca también C11. El clon al que los Criollos se parecen menos en semilla es el Forastero EET-16. Los clones de alto rendimiento permanecen cercanos.

En 3D (**Gráfico 21-3**) destacan C3, C9 por un lado, C6 por otro, EET-116 en la parte superior y C11 en la parte inferior.

De los clones de alto rendimiento (**Gráfico 22-3**) los más similares en semilla son C8 y C25, C2 y C26, y C16 y C19. En general se mantienen cercanos. Ninguno se aproxima a EET-116 (Forastero), y se aproximan más a CCN-51 y a Nacional EET-103. Se distancian un poco de Trinitario EET-111.

Tabla 14-3: Variabilidad explicada por las coordenadas principales de la configuración de semilla.

Coordenada Principal - Dimensión q (q=1, …, 29)	Autovalor	Variabilidad geométrica explicada por la coordenada q	Fracción de variabilidad explicada por la coordenada q	Porcentaje de variabilidad explicada hasta los q primeros ejes
1	46.4	1.5	0.12	12%
2	41.2	1.4	0.10	22%
3	21.1	0.7	0.05	27%
4	17.6	0.6	0.04	32%
5	16.2	0.5	0.04	36%
6	13.7	0.5	0.03	39%
7	12.5	0.4	0.03	43%
8	12.3	0.4	0.03	46%
9	12.1	0.4	0.03	49%
10	11.8	0.4	0.03	52%
11	11.6	0.4	0.03	55%
12	11.6	0.4	0.03	57%
13	11.6	0.4	0.03	60%
14	11.5	0.4	0.03	63%
15	11.5	0.4	0.03	66%
16	11.5	0.4	0.03	69%
17	11.4	0.4	0.03	72%
18	11.4	0.4	0.03	75%
19	11.4	0.4	0.03	78%
20	11.4	0.4	0.03	80%
21	11.3	0.4	0.03	83%
22	11.2	0.4	0.03	86%
23	11.1	0.4	0.03	89%
24	10.9	0.4	0.03	92%
25	10.6	0.4	0.03	94%
26	10.1	0.3	0.03	97%
27	9.8	0.3	0.02	99%
28	2.4	0.1	0.01	100%
29	0.0	0.0	0.00	100%
Variabilidad geométrica		13.2	1.00	

Realizado por: Gabriela J. Obregón O. 2018.

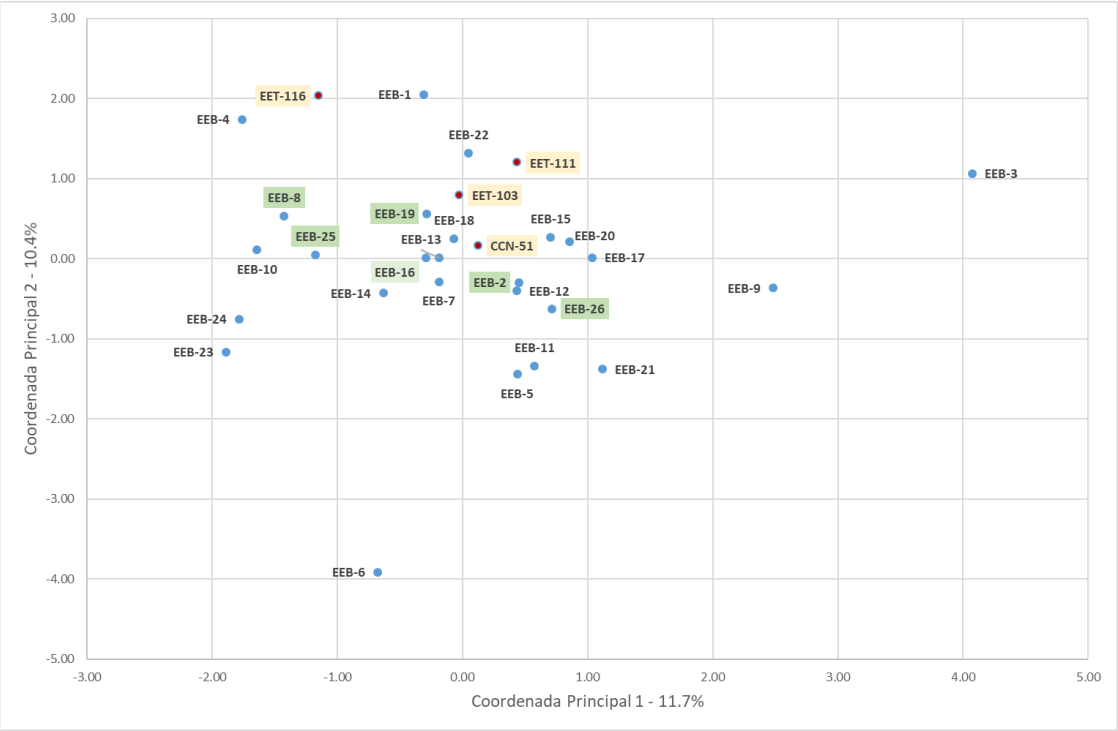

Gráfico 18-3: Coordenadas principales 1 y 2 de similitudes de semilla. Explican el 22.1% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

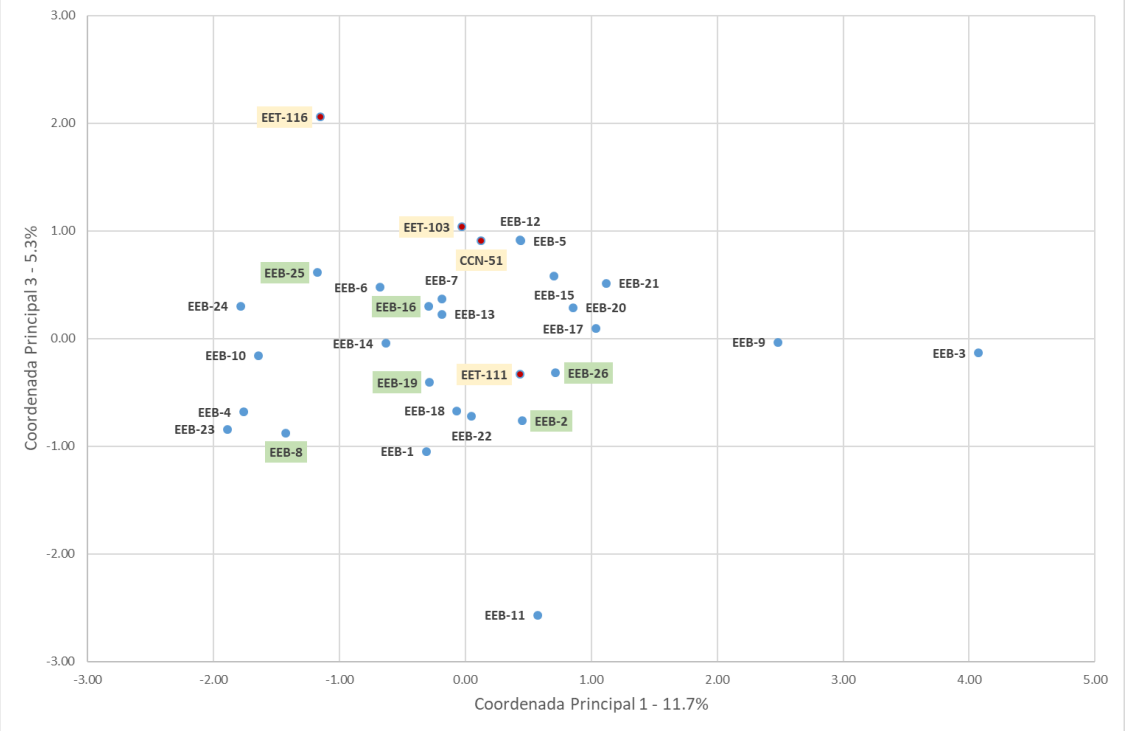

Gráfico 19-3: Coordenadas principales 1 y 3 de similitudes de semilla. Explican el 17% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

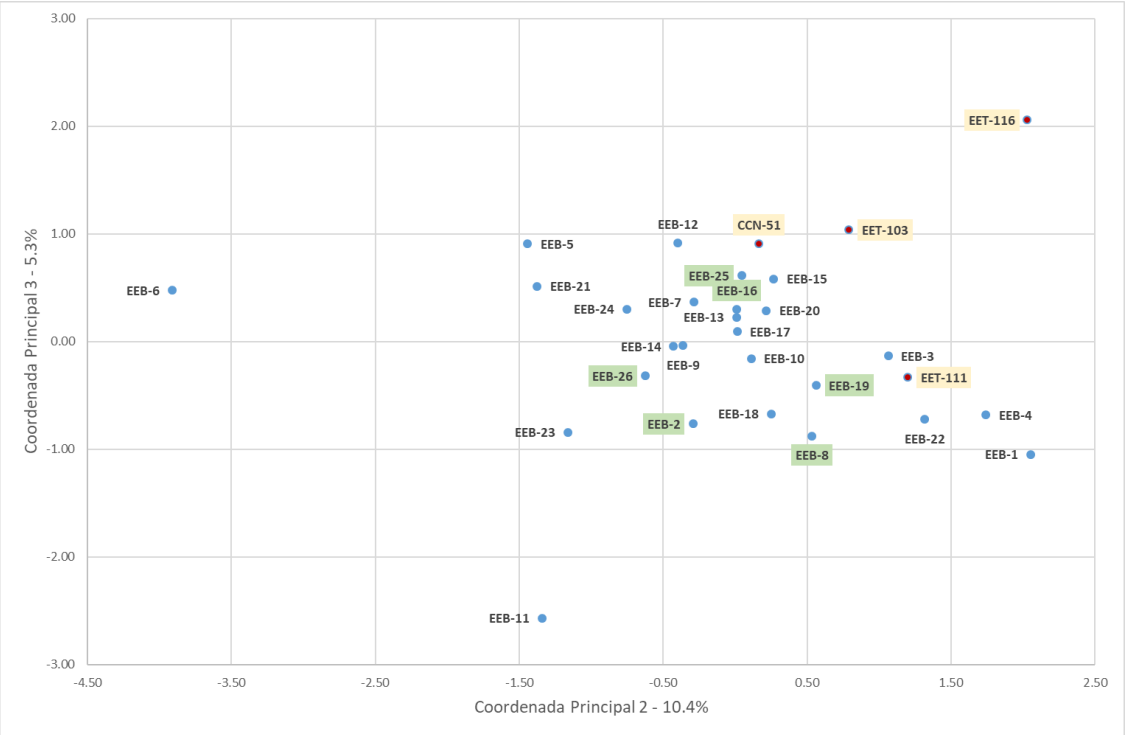

Gráfico 20-3: Coordenadas principales 2 y 3 de similitudes de semilla. Explican el 15.7% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

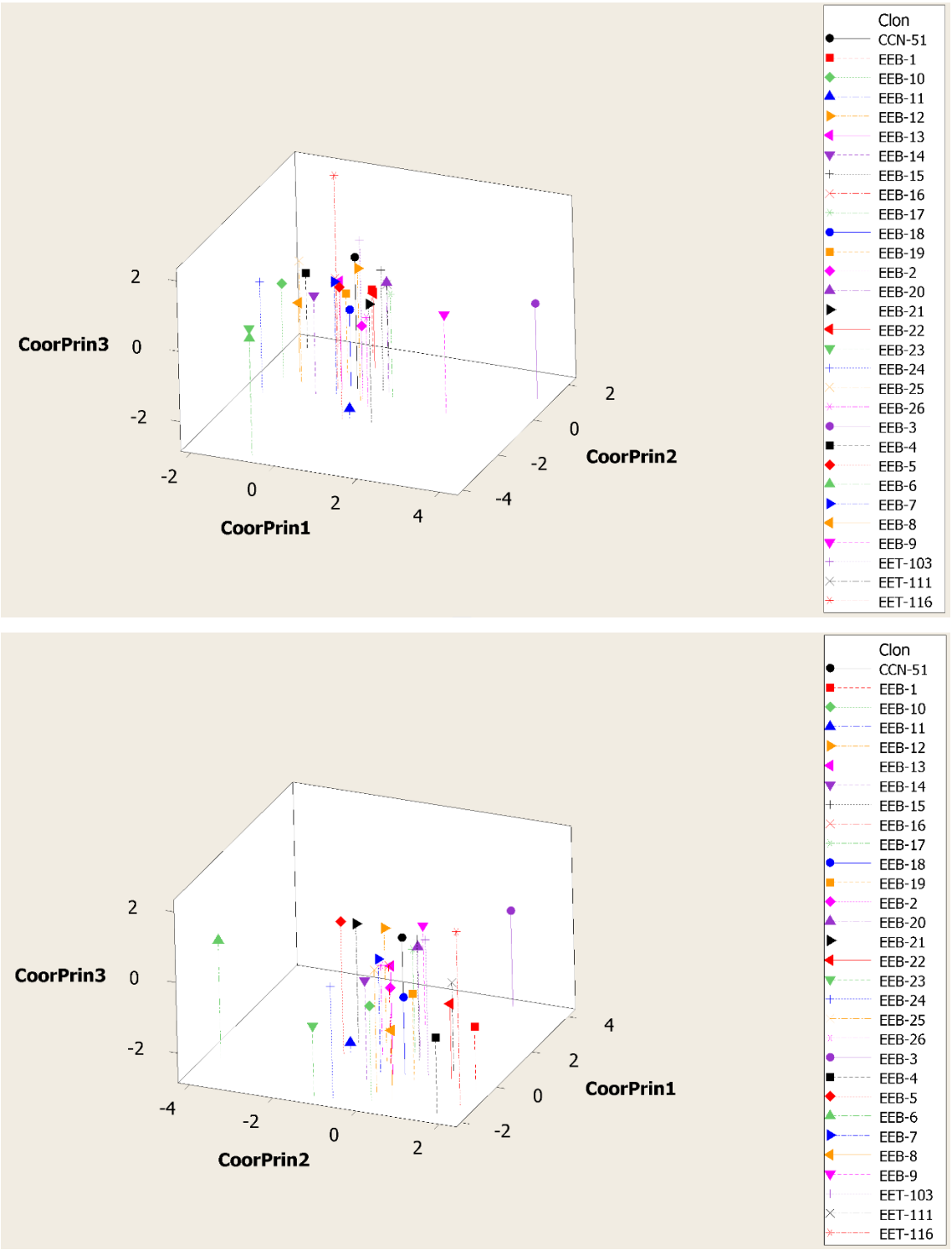

Gráfico 21-3: Visualización en 3D con las 3 primeras Coordenadas Principales de datos de semilla, desde diferentes perspectivas.

Realizado por: Gabriela J. Obregón O. 2018.

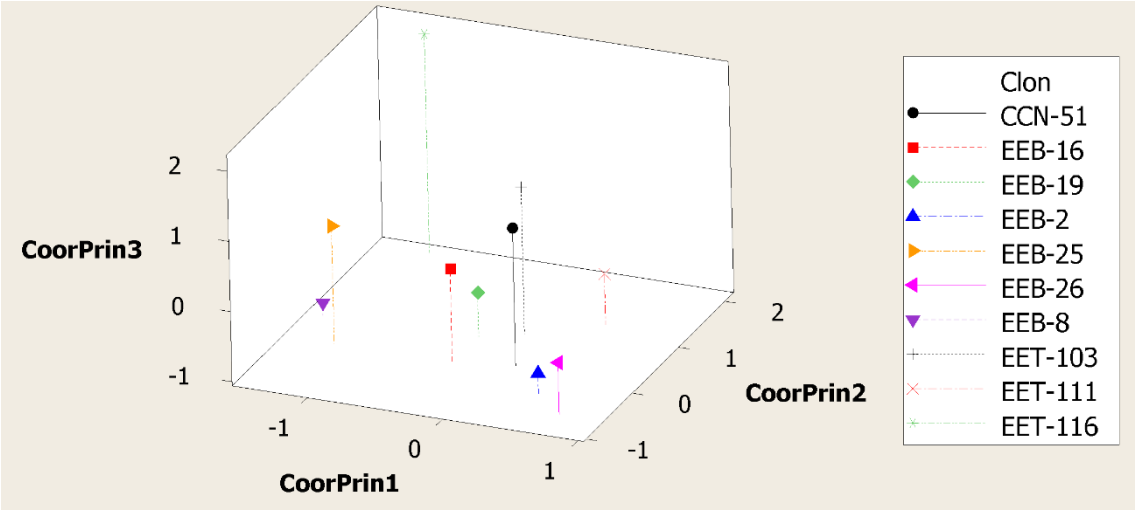

Gráfico 22-3: Gráfica en 3D con enfoque en los 6 de tipo Criollo de más alto rendimiento, en comparación con los testigos.

Realizado por: Gabriela J. Obregón O. 2018.

3.6.4 Configuración de Hoja

Las 3 primeras coordenadas principales explican el 32% de la variabilidad.

Destacan en el **Gráfico 23-3** C4 en el lado izquierdo, y C23, C17 y C5 en el lado derecho. Los clones testigo no se mezclan con los de tipo Criollo tanto en éste como en el **Gráfico 24-3** y **25-3**, destacando CCN-51, EET-111 y EET-103.

En 3D (**Gráfico 26-3**) el que más se aleja es CCN-51, y los demás clones testigo están cercanos a éste y no se ubican entre los de tipo Criollo.

Siguiendo los de alto rendimiento, un grupo separado podría ser C2 y C26. Se encuentran cercanos EET-111, EET-103 y C8.

Tabla 15-3: Variabilidad explicada por las coordenadas principales de la configuración de hoja.

Coordenada Principal - Dimensión q (q=1, …, 28)	Autovalor	Variabilidad geométrica explicada por la coordenada q	Fracción de variabilidad explicada por la coordenada q	Porcentaje de variabilidad explicada hasta los q primeros ejes
1	69.0504	2.302	0.15	15%
2	49.0649	1.635	0.11	26%
3	24.7722	0.826	0.05	32%
4	20.9268	0.698	0.05	36%
5	20.2938	0.676	0.05	41%
6	17.5150	0.584	0.04	45%
7	14.7632	0.492	0.03	48%
8	14.1198	0.471	0.03	51%
9	13.7571	0.459	0.03	54%
10	13.2965	0.443	0.03	57%
11	13.1105	0.437	0.03	60%
12	12.5879	0.420	0.03	63%
13	12.2555	0.409	0.03	66%
14	12.0520	0.402	0.03	68%
15	11.9678	0.399	0.03	71%
16	11.8872	0.396	0.03	74%
17	11.6816	0.389	0.03	76%
18	11.6335	0.388	0.03	79%
19	11.4571	0.382	0.03	81%
20	11.3685	0.379	0.03	84%
21	11.0237	0.367	0.02	86%
22	10.7836	0.359	0.02	89%
23	10.5711	0.352	0.02	91%
24	10.0436	0.335	0.02	93%
25	9.6316	0.321	0.02	95%
26	8.8173	0.294	0.02	97%
27	6.3038	0.210	0.01	99%
28	5.9320	0.198	0.01	100%
Variabilidad geométrica		15.022	1.00	

Realizado por: Gabriela J. Obregón O. 2018.

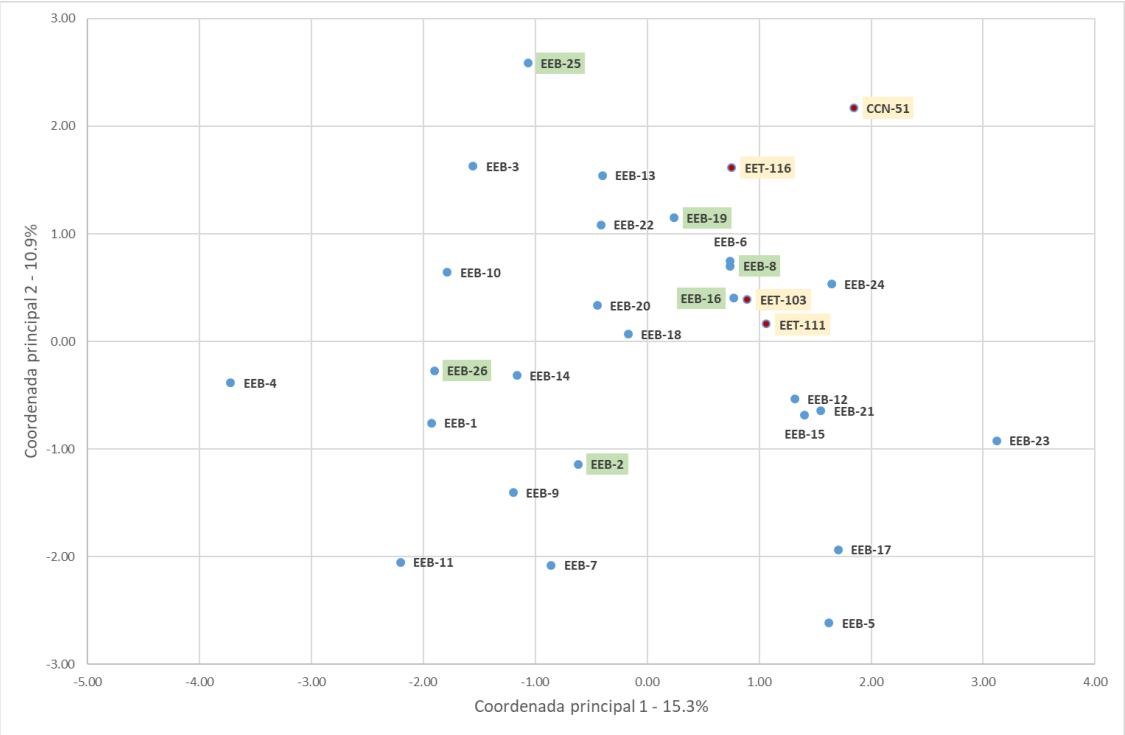

Gráfico 23-3: Coordenadas principales 1 y 2 de similitudes de hoja. Explican el 26.2% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

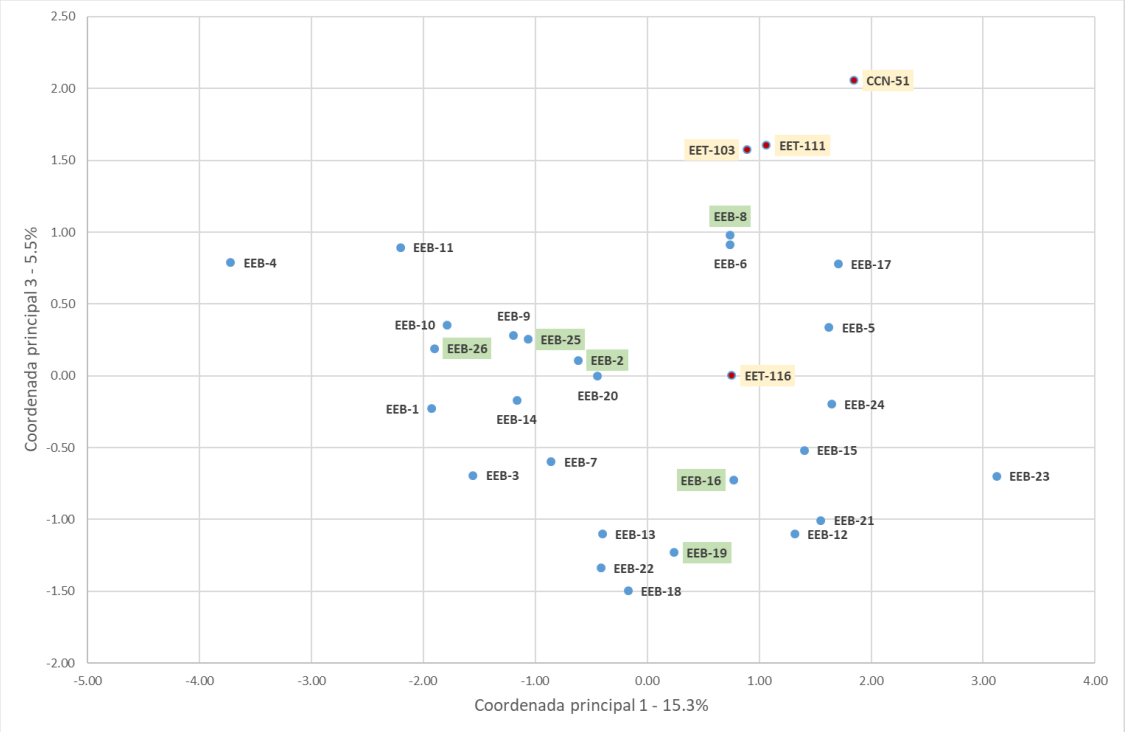

Gráfico 24-3: Coordenadas principales 1 y 3 de similitudes de hoja. Explican el 20.8% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

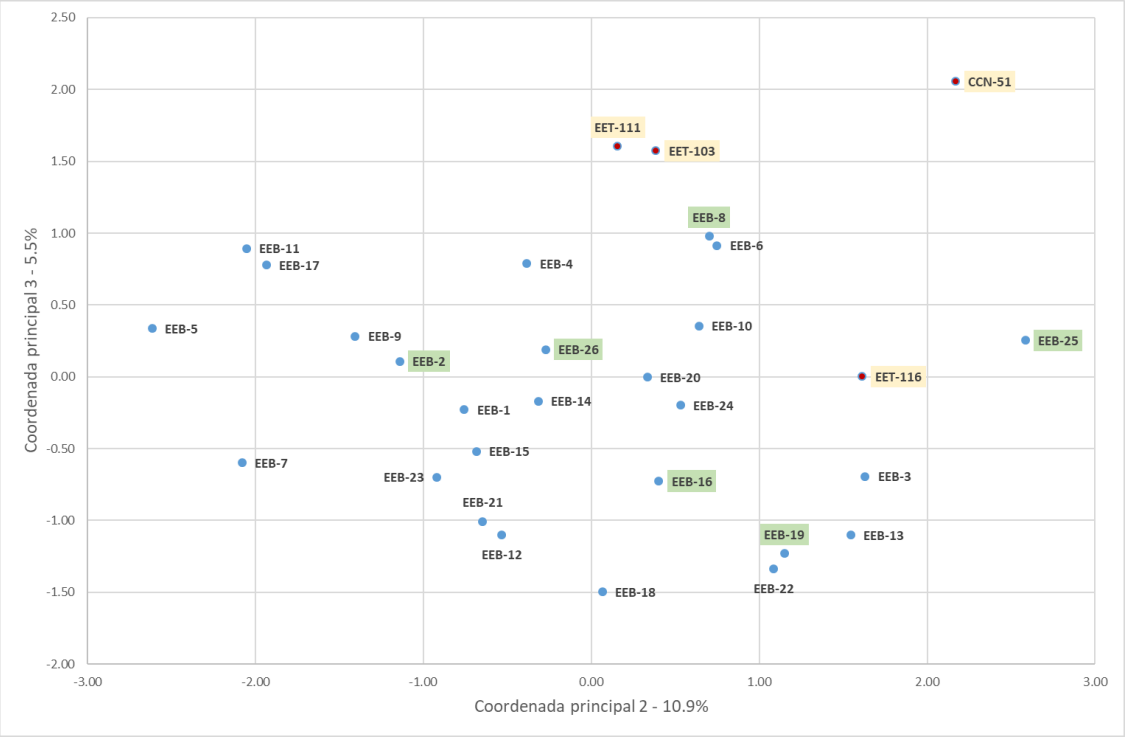

Gráfico 25-3: Coordenadas principales 2 y 3 de similitudes de hoja. Explican el 16.4% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

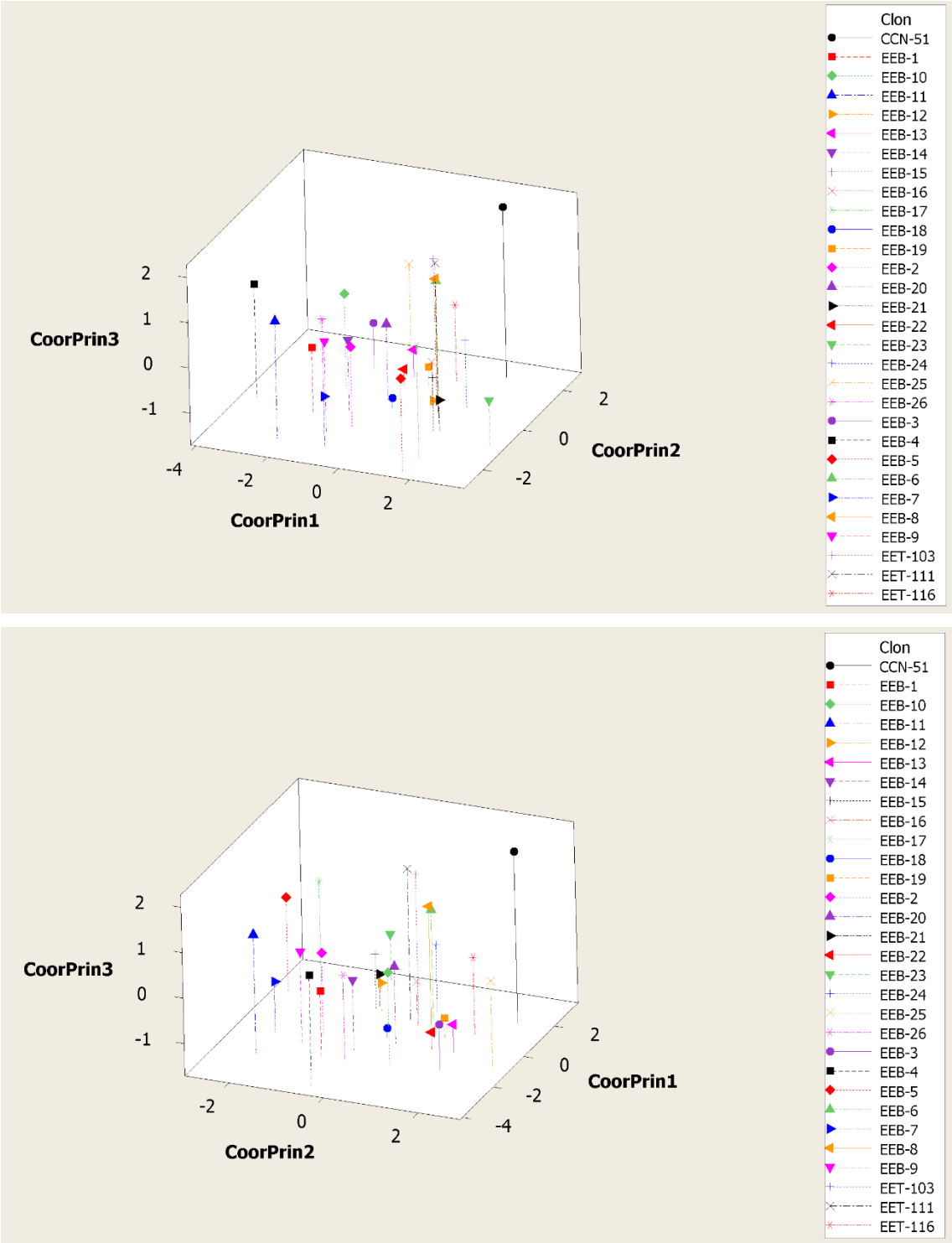

Gráfico 26-3: Visualización en 3D con las 3 primeras Coordenadas Principales de datos de hoja, desde diferentes perspectivas.

Realizado por: Gabriela J. Obregón O. 2018.

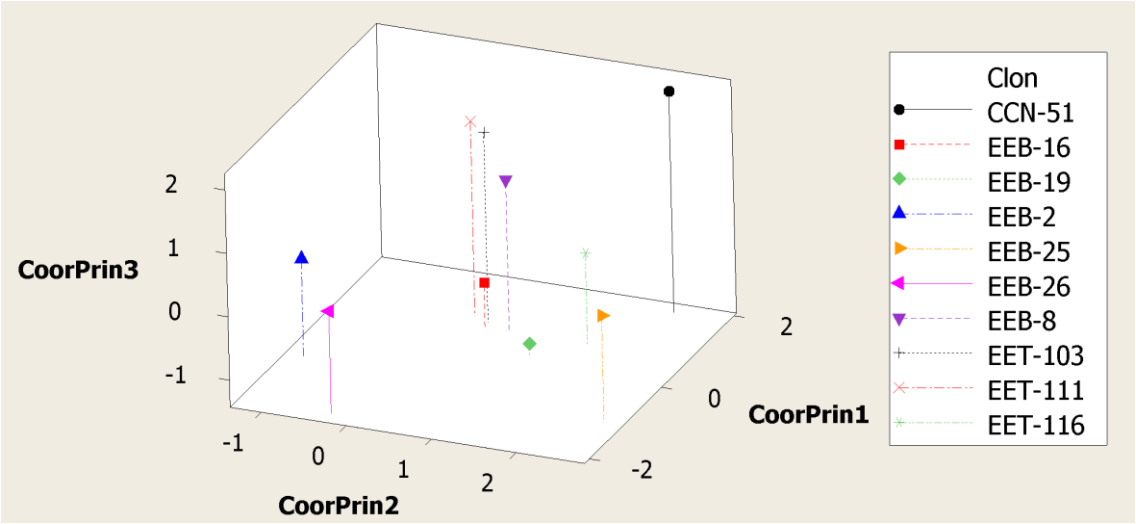

Gráfico 27-3: Gráfica en 3D con enfoque en los 6 de tipo Criollo de más alto rendimiento, en comparación con los testigos.

Realizado por: Gabriela J. Obregón O. 2018.

3.6.5 Configuración de Flor

El porcentaje de variabilidad explicado por las 3 primeras coordenadas es de 23%.

Tabla 16-3: Variabilidad explicada por las coordenadas principales de la configuración de flor.

Coordenada Principal - Dimensión q (q=1, …, 30)	Autovalor	Variabilidad geométrica explicada por la coordenada q	Fracción de variabilidad explicada por la coordenada q	Porcentaje de variabilidad explicada hasta los q primeros ejes
1	30.0000	1.000	0.08	8%
2	30.0000	1.000	0.08	15%
3	30.0000	1.000	0.08	23%
4	30.0000	1.000	0.08	31%
5	30.0000	1.000	0.08	38%
6	30.0000	1.000	0.08	46%
7	30.0000	1.000	0.08	54%
8	30.0000	1.000	0.08	62%
9	30.0000	1.000	0.08	69%
10	30.0000	1.000	0.08	77%
11	30.0000	1.000	0.08	85%
12	30.0000	1.000	0.08	92%
13	30.0000	1.000	0.08	100%
Variabilidad geométrica		13.000	1.00	

Realizado por: Gabriela J. Obregón O. 2018.

Claramente destaca la flor de EET-111, además de C17 en el **Gráfico 28-3**. En los siguientes gráficos destacan además C1, C8, C19 y C4.

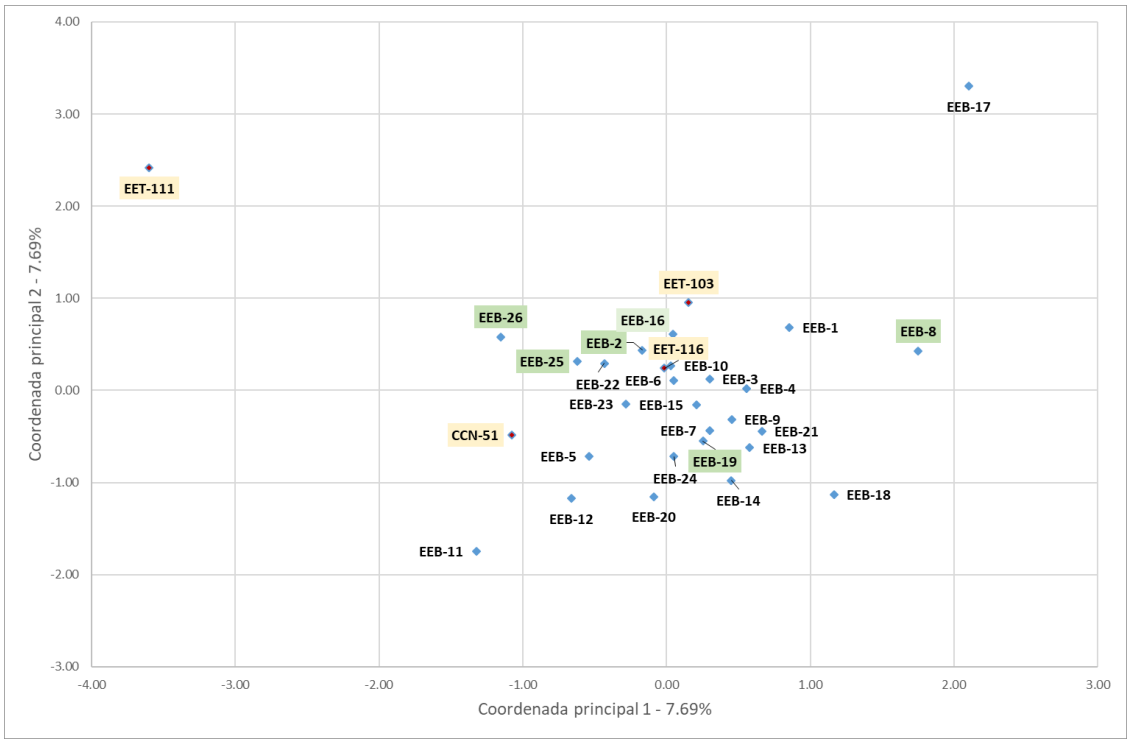

Gráfico 28-3: Coordenadas principales 1 y 2 de similitudes de flor. Explican el 15.4% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

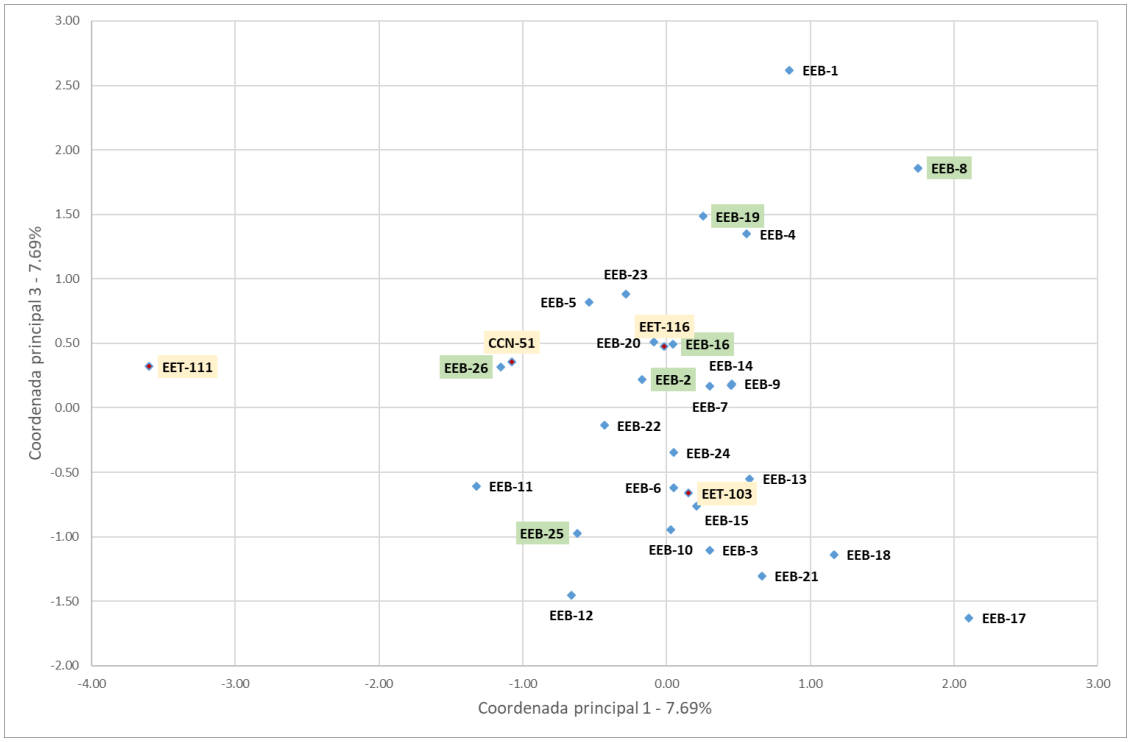

Gráfico 29-3: Coordenadas principales 1 y 3 de similitudes de flor. Explican el 15.4% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

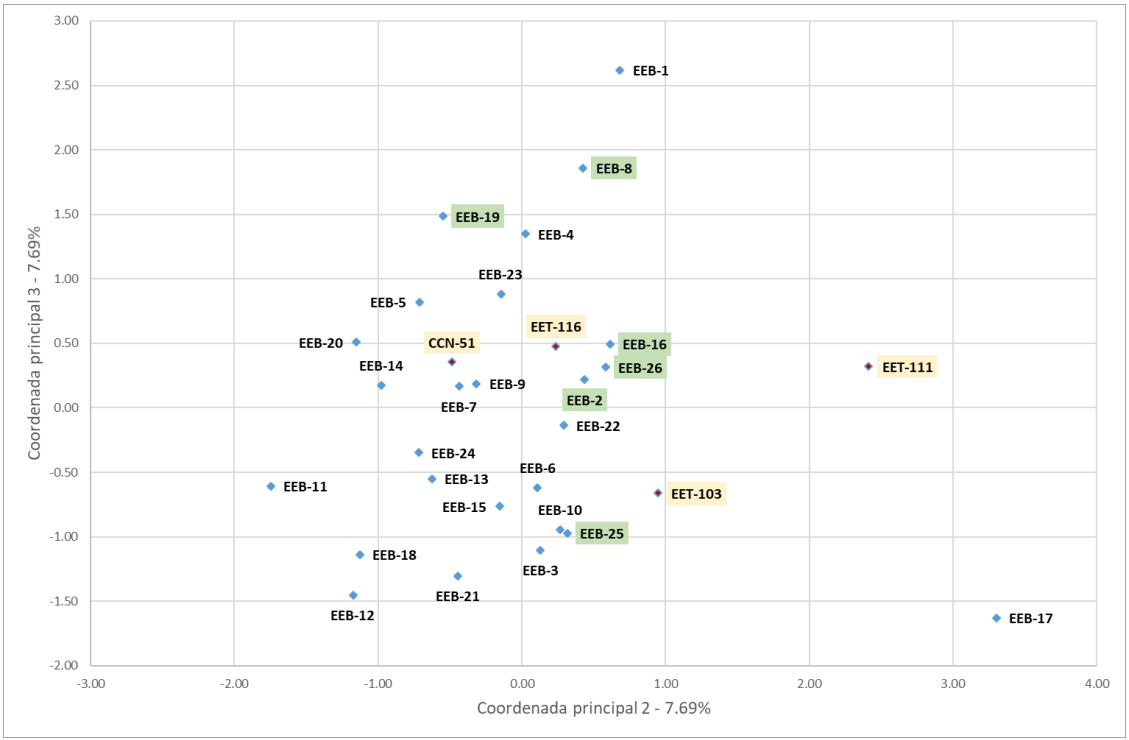

Gráfico 30-3: Coordenadas principales 2 y 3 de similitudes de flor. Explican el 15.4% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

En 3D (**Gráfico 31-3**) se aprecia que se alejan C17 y EET-111. En la parte superior destacan C1, C8, C4 y C19.

En el enfoque en los de alto rendimiento (**Gráfico 32-3**), están cercanos todos excepto C8, y se asemejan más a EET-103. Ninguno se parece a EET-111.

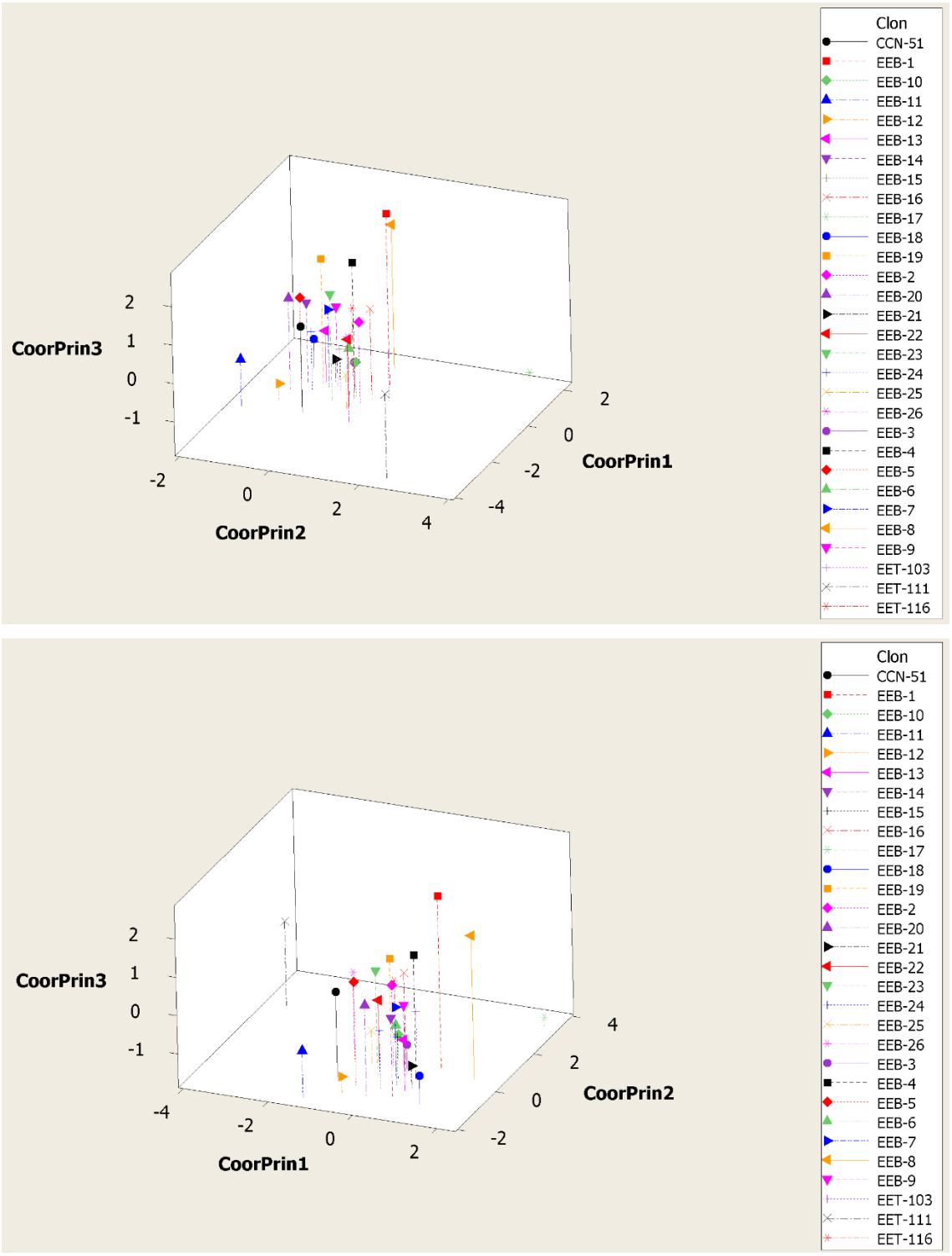

Gráfico 31-3: Visualización en 3D con las 3 primeras Coordenadas Principales de datos de flor, desde diferentes perspectivas.

Realizado por: Gabriela J. Obregón O. 2018.

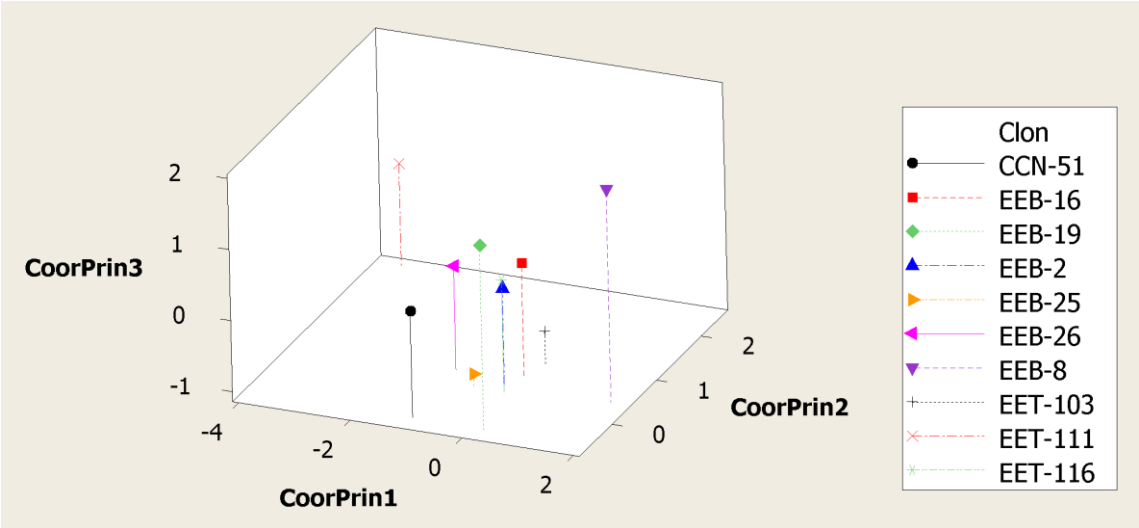

Gráfico 32-3: Gráfica en 3D con enfoque en los 6 de tipo Criollo de más alto rendimiento, en comparación con los testigos.

Realizado por: Gabriela J. Obregón O. 2018.

3.7 Análisis Procrustes Generalizado

3.7.1 Consenso entre configuraciones morfológicas

Las 3 primeras componentes principales del consenso morfológico reflejan el 21% de la variabilidad, por lo que en 3 dimensiones la representación no es muy adecuada.

La variabilidad total en este caso se reparte en consenso + residual. La suma total puede descomponerse en *por clones* o *por configuraciones*. Cada consenso en cada clon es la concordancia entre las 4 caracterizaciones morfológicas: mazorca, semilla, hoja y flor.

En la **Tabla 18-3** se aprecia que C4 tuvo el residual más alto y el porcentaje de consenso más bajo (71.6%), es decir que en éste es donde las 4 configuraciones difieren más. En cambio, C1 tuvo el residual más bajo y el porcentaje de consenso más alto (95.8%), es decir que las 4 configuraciones tuvieron mayor concordancia en este clon.

En la **Tabla 19-3** se observa que la configuración de flor tuvo el residual más alto y el porcentaje de consenso más bajo (53.6%), es decir que difiere más del consenso, en lo cual quizá influyó que, debido a las limitaciones, no se pudieron utilizar las flores como unidad de análisis directamente, y en su lugar se utilizaron los promedios por clon de cada variable. En cambio, las configuraciones de mazorca y semilla obtuvieron un menor residuo y mayor porcentaje de consenso (91.2% y 91.1% respectivamente), por lo que éstas no difieren mucho del consenso.

Tabla 17-3: Autovalores de los componentes principales del consenso morfológico.

Lambda	Valor	Proporción	Prop. Acumulada
1	0.07	0.08	0.08
2	0.06	0.07	0.15
3	0.05	0.06	0.21
4	0.05	0.05	0.27
5	0.04	0.05	0.32
6	0.04	0.05	0.37
7	0.04	0.05	0.42
8	0.04	0.05	0.46
9	0.04	0.04	0.51
10	0.04	0.04	0.55
11	0.04	0.04	0.59
12	0.04	0.04	0.63
13	0.03	0.04	0.67
14	0.03	0.03	0.70
15	0.02	0.03	0.73
16	0.02	0.03	0.76
17	0.02	0.03	0.78
18	0.02	0.02	0.80
19	0.02	0.02	0.83
20	0.02	0.02	0.85
21	0.02	0.02	0.87
22	0.02	0.02	0.89
23	0.02	0.02	0.91
24	0.02	0.02	0.93
25	0.02	0.02	0.95
26	0.02	0.02	0.97
27	0.01	0.02	0.99
28	0.01	0.02	1.00
29	0.00	0.00	1.00

Realizado por: Gabriela J. Obregón O. 2018.

Tanto en la **Tabla 18-3** como en la **Tabla 19-3**, se puede conocer el valor de consenso final, que es de 83.7%, esto significa que existe un alto consenso entre los 4 tipos de caracterizaciones morfológicas. En los **Gráficos 33-3**, **34-3** y **35-3** se representan los primeros componentes principales de este consenso morfológico. Las etiquetas de los testigos aparecen en color rojo, las de los de alto rendimiento en color verde, y las de bajo rendimiento en color amarillo.

En el **Gráfico 33-3** se aprecia que destacan del grupo central CCN-51 en la parte superior, C1 y C4 en la parte derecha, C5 y C6 en la parte izquierda, y C9 en la parte inferior.

En el **Gráfico 34-3** destacan C3 y C9 en la parte superior. De igual manera en el **Gráfico 35-3**.

En general ninguno de los Criollos de alto rendimiento se aleja del grupo.

En 3D (**Gráfico 36-3**) se alejan del grupo central: C3, CCN-51, C5, C6, C9, C4 y C1.

Tabla 18-3: Cuadro de Análisis de Varianza. Suma de cuadrados por Clon.

Clon	Consenso	Residuo	Total	%Consenso
C1	0.158	0.008	0.165	95.8%
C2	0.112	0.018	0.129	86.8%
C3	0.138	0.028	0.165	83.6%
C4	0.116	0.046	0.162	71.6%
C5	0.134	0.014	0.148	90.5%
C6	0.125	0.032	0.157	79.6%
C7	0.102	0.025	0.127	80.3%
C8	0.101	0.019	0.12	84.2%
C9	0.15	0.009	0.158	94.9%
C10	0.124	0.015	0.139	89.2%
C11	0.11	0.029	0.139	79.1%
C12	0.087	0.023	0.11	79.1%
C13	0.092	0.02	0.112	82.1%
C14	0.111	0.015	0.126	88.1%
C15	0.115	0.014	0.129	89.1%
C16	0.107	0.022	0.129	82.9%
C17	0.131	0.014	0.145	90.3%
C18	0.086	0.023	0.109	78.9%
C19	0.095	0.02	0.114	83.3%
C20	0.085	0.025	0.11	77.3%
C21	0.107	0.023	0.13	82.3%
C22	0.087	0.024	0.111	78.4%
C23	0.108	0.032	0.14	77.1%
C24	0.097	0.023	0.119	81.5%
C25	0.09	0.026	0.116	77.6%
C26	0.106	0.018	0.124	85.5%
CCN51	0.123	0.037	0.159	77.4%
EET116	0.109	0.026	0.135	80.7%
EET111	0.133	0.012	0.145	91.7%
EET103	0.109	0.017	0.126	86.5%
Total	**3.346**	**0.654**	**4**	**83.7%**

Realizado por: Gabriela J. Obregón O. 2018.

Tabla 19-3: Cuadro de Análisis de Varianza. Suma de cuadrados por Configuración.

Configuración	Consenso	Residuo	Total	%Consenso
Mazorca	1.001	0.097	1.098	91.2%
Semilla	0.998	0.098	1.096	91.1%
Hoja	0.949	0.118	1.067	88.9%
Flor	0.397	0.342	0.74	53.6%
Total	**3.346**	**0.654**	**4**	**83.7%**

Realizado por: Gabriela J. Obregón O. 2018.

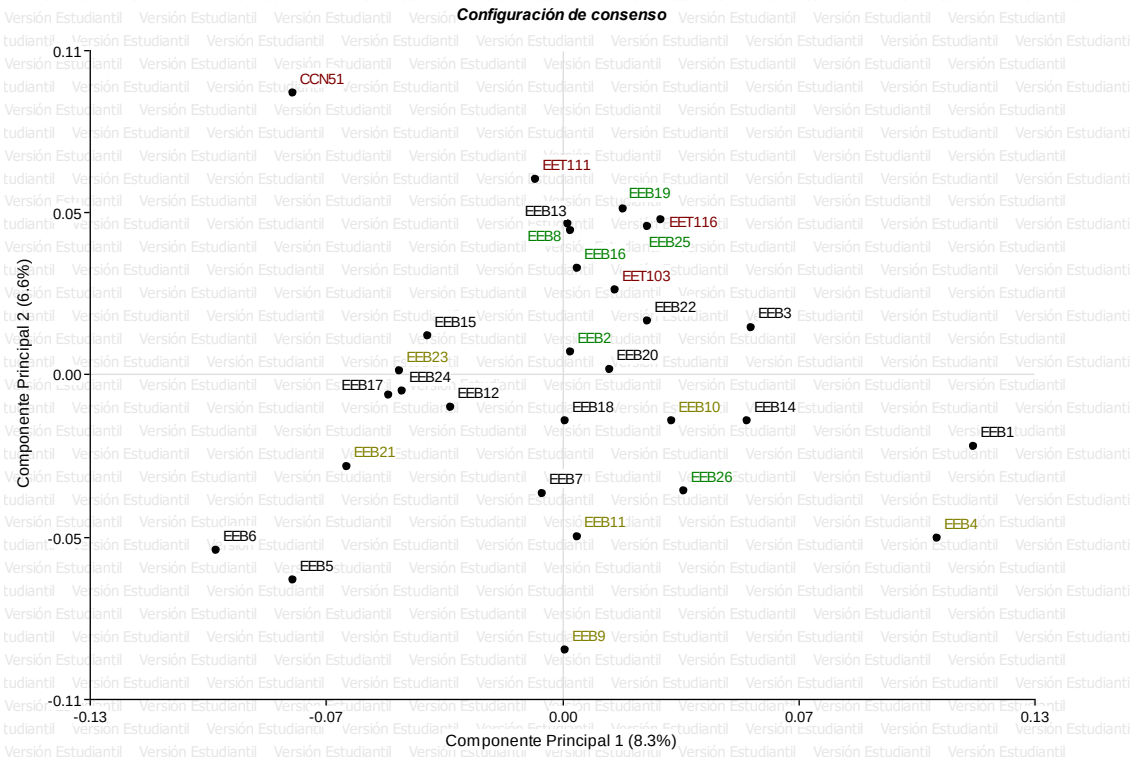

Gráfico 33-3: Componentes principales 1 y 2 del consenso morfológico. Explican el 14.9% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

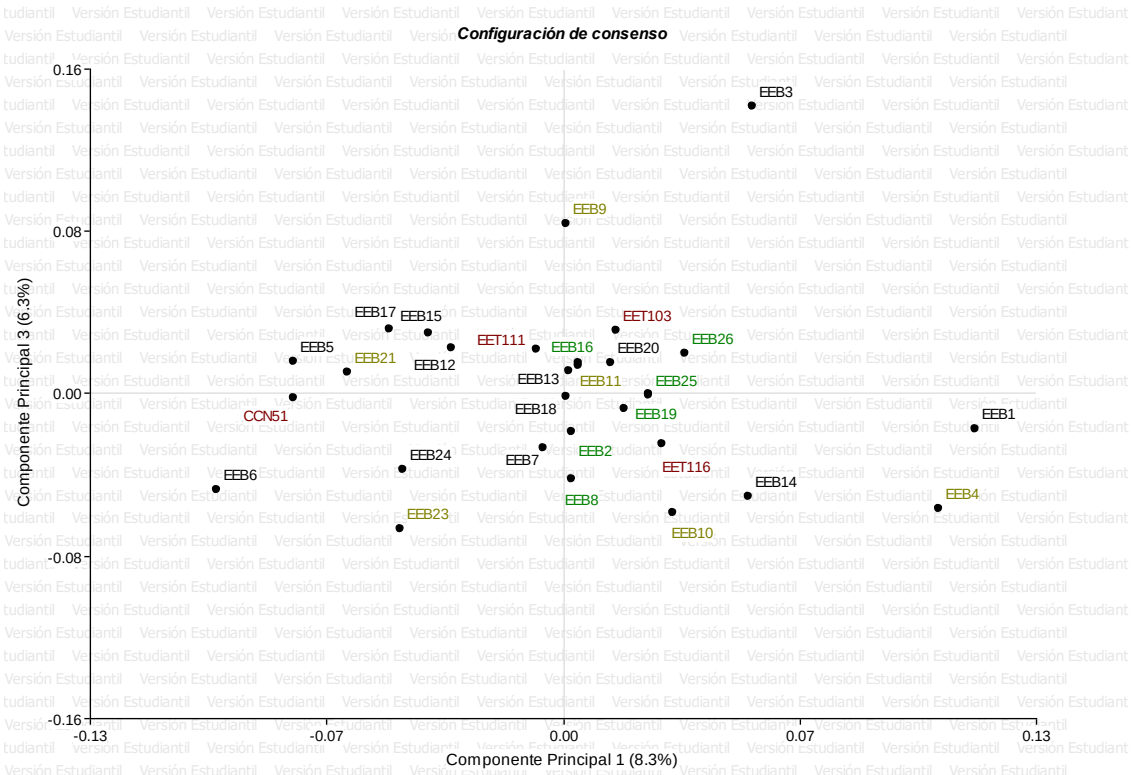

Gráfico 34-3: Componentes principales 1 y 3 del consenso morfológico. Explican el 14.6% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

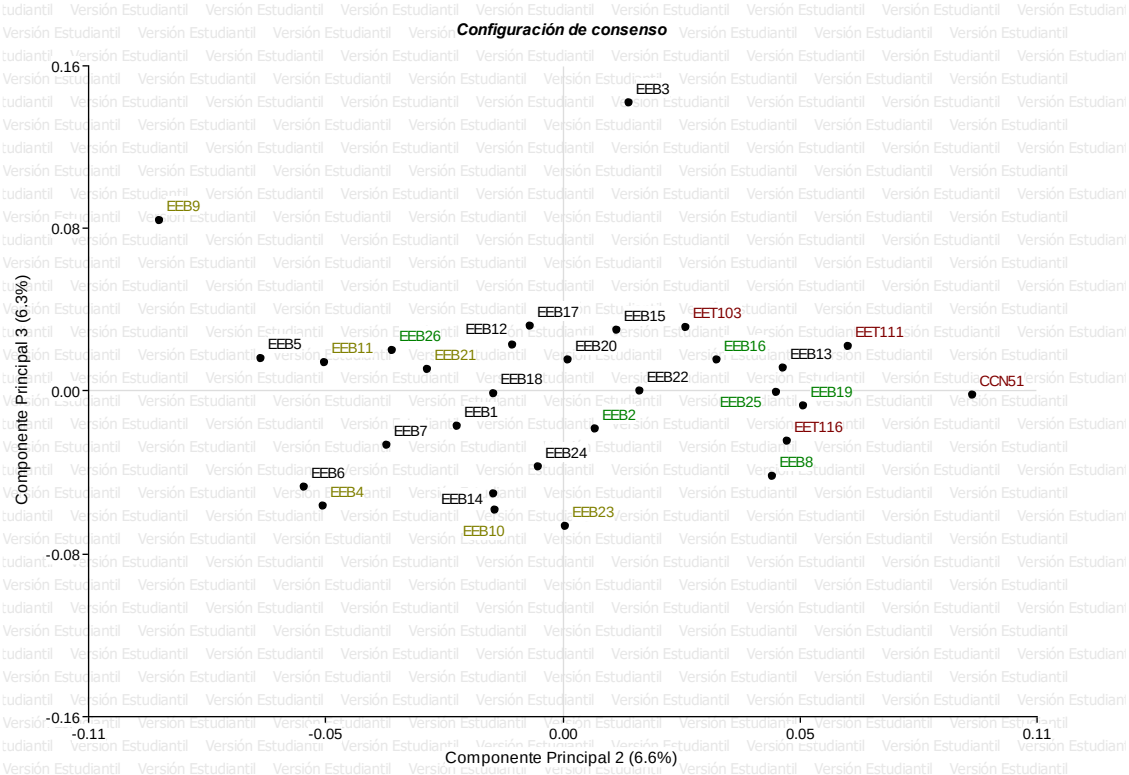

Gráfico 35-3: Componentes principales 2 y 3 del consenso morfológico. Explican el 12.9% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

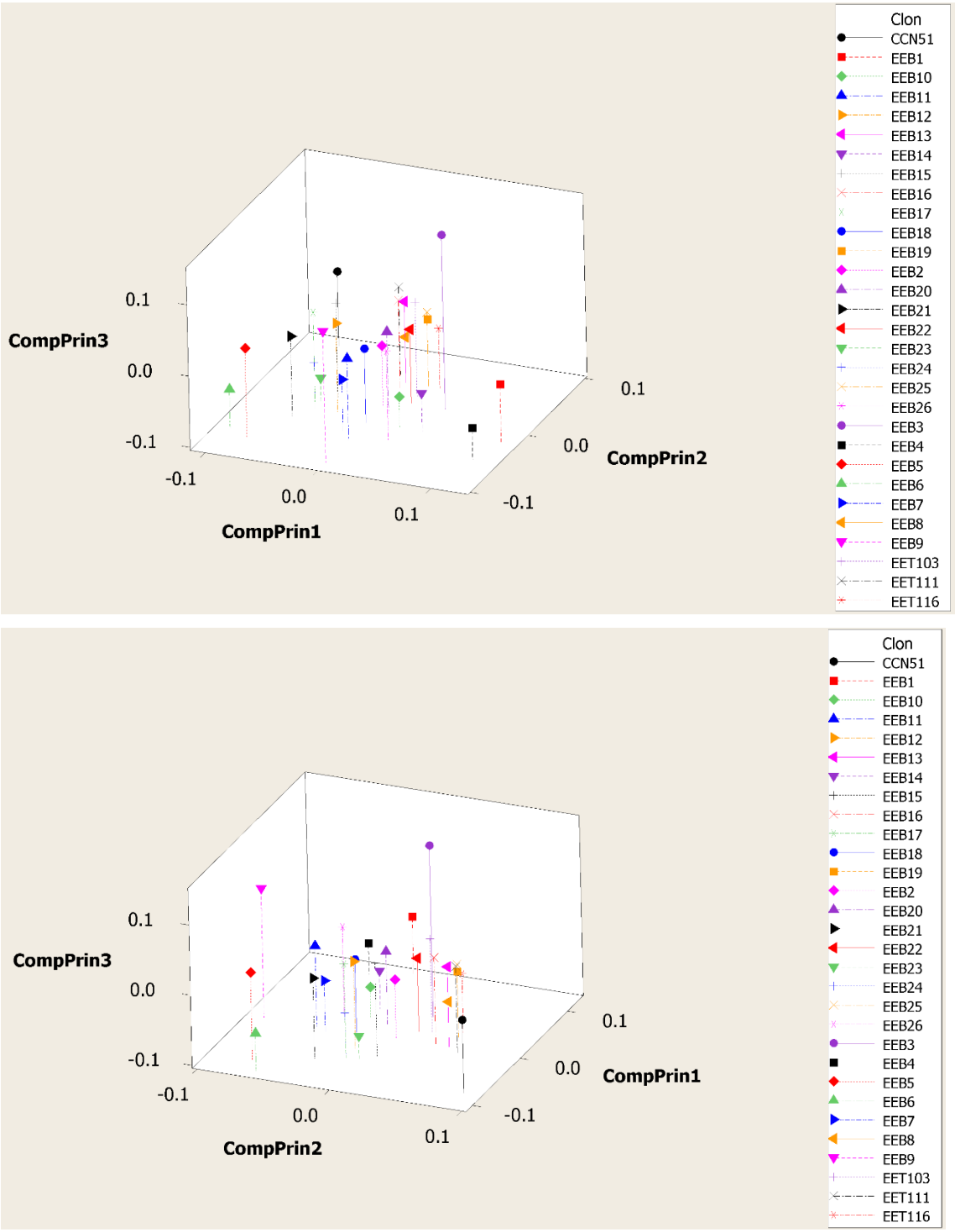

Gráfico 36-3: Visualización en 3D con los 3 primeros componentes principales del consenso, desde diferentes perspectivas.

Realizado por: Gabriela J. Obregón O. 2018.

En el seguimiento a los de alto rendimiento (**Gráfico 37-3**), y en comparación con los de más bajo rendimiento, en características morfológicas se observan en general separados de éstos y más cercanos a los testigos.

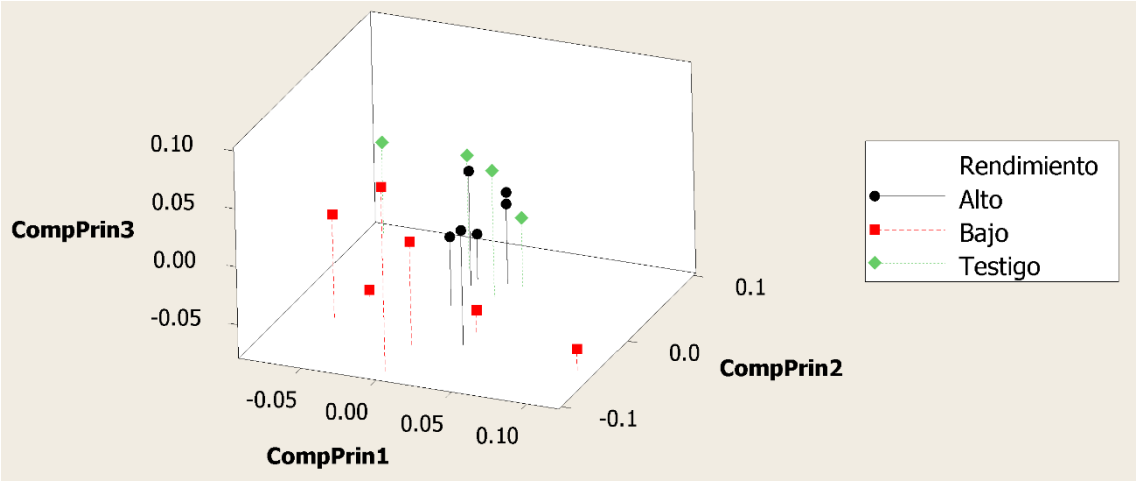

Gráfico 37-3: Posición de Criollos de alto rendimiento con respecto a los de bajo rendimiento y testigos.

Realizado por: Gabriela J. Obregón O. 2018.

3.7.2 Consenso entre configuración molecular y el consenso morfológico

Tabla 20-3: Autovalores de los componentes principales del consenso morfológico-molecular.

Lambda	Valor	Proporción	Prop. Acum
1	0.08	0.09	0.09
2	0.07	0.07	0.16
3	0.06	0.06	0.22
4	0.06	0.06	0.28
5	0.05	0.05	0.34
6	0.05	0.05	0.39
7	0.05	0.05	0.44
8	0.04	0.04	0.48
9	0.04	0.04	0.52
10	0.04	0.04	0.57
11	0.04	0.04	0.61
12	0.03	0.04	0.64
13	0.03	0.03	0.68
14	0.03	0.03	0.71
15	0.03	0.03	0.74
16	0.03	0.03	0.76
17	0.03	0.03	0.79
18	0.02	0.03	0.81
19	0.02	0.02	0.84
20	0.02	0.02	0.86
21	0.02	0.02	0.88
22	0.02	0.02	0.90
23	0.02	0.02	0.92
24	0.02	0.02	0.94
25	0.02	0.02	0.95
26	0.01	0.02	0.97
27	0.01	0.02	0.98
28	0.01	0.01	1.00
29	0.01	0.01	1.00

Realizado por: Gabriela J. Obregón O. 2018.

Las 3 primeras componentes principales explican el 22% de la variabilidad, lo cual no es una representación muy adecuada.

En la **Tabla 21-3** se aprecia que C16 tuvo el residual más alto y el porcentaje de consenso más bajo (89.2%), es decir que en éste es donde las 2 configuraciones difieren más. Le siguen en residual alto C21 (90.1% consenso) y C23 (92.1% consenso), de los que se concluye lo mismo. En cambio, EET-103 tuvo el residual más bajo y el porcentaje de consenso más alto (98.3%), es decir que las 2 configuraciones tuvieron mayor consenso en este clon.

Tabla 21-3: Cuadro de Análisis de Varianza. Suma de cuadrados por Clon.

Clon	Consenso	Residuo	Total	%Consenso
C1	0.088	0.003	0.09	97.2%
C2	0.064	0.004	0.068	93.8%
C3	0.074	0.003	0.077	95.9%
C4	0.058	0.006	0.064	90.4%
C5	0.064	0.004	0.068	93.9%
C6	0.06	0.004	0.064	93.2%
C7	0.055	0.006	0.061	89.6%
C8	0.06	0.003	0.063	95.9%
C9	0.078	0.003	0.082	96.2%
C10	0.059	0.004	0.063	93.7%
C11	0.063	0.004	0.066	94.7%
C12	0.053	0.003	0.055	95.4%
C13	0.054	0.004	0.058	93.4%
C14	0.061	0.002	0.063	96.9%
C15	0.056	0.004	0.06	93.8%
C16	0.059	0.007	0.066	89.2%
C17	0.073	0.003	0.077	95.7%
C18	0.05	0.002	0.053	95.6%
C19	0.046	0.003	0.049	93.7%
C20	0.05	0.004	0.054	93.2%
C21	0.062	0.007	0.068	90.1%
C22	0.052	0.002	0.055	95.6%
C23	0.079	0.007	0.086	92.1%
C24	0.054	0.003	0.056	95.2%
C25	0.054	0.004	0.058	93.1%
C26	0.072	0.004	0.076	95.1%
CCN51	0.075	0.004	0.079	95.2%
EET116	0.08	0.006	0.086	93.3%
EET111	0.071	0.002	0.073	97.6%
EET103	0.06	0.001	0.061	98.3%
Total	**1.886**	**0.114**	**2**	**94.3%**

Realizado por: Gabriela J. Obregón O. 2018.

En la **Tabla 22-3** se aprecia que tanto el consenso morfológico como la configuración molecular tienen un porcentaje de consenso de 94.3%, que se considera alto, por lo que éstas no difieren mucho del consenso.

Tabla 22-3: Cuadro de Análisis de Varianza. Suma de cuadrados por Configuración.

Configuración	Consenso	Residuo	Total	%Consenso
Consenso morfológico	0.943	0.057	1	94.3%
Molecular	0.943	0.057	1	94.3%
Total	**1.886**	**0.114**	**2**	**94.3%**

Realizado por: Gabriela J. Obregón O. 2018.

Tanto en la **Tabla 21-3** como en la **Tabla 22-3**, se puede conocer el valor de consenso final, que es de 94.3%, esto significa que existe un alto consenso entre las 2 configuraciones. En los **Gráficos 38-3**, **39-3** y **40-3** se representan las primeras componentes principales.

En el **Gráfico 38-3** los clones C23, EET-116 y CCN-51 se distancian de los demás hacia la derecha. Los testigos EET-111 y EET-103 permanecen cercanos y los Criollos de alto rendimiento están próximos a estos. Se diría que los de tipo Criollo forman un grupo. En las demás gráficas destacan C3, C9, C17 y C2.

En 3D (**Gráfico 41-3**) los posibles grupos son: CCN-51, EET-116 y C23; C9, C21, C16, C3 y C17; el resto se diría que forma otro grupo.

Con respecto a los Criollos de alto rendimiento (**Gráfico 42-3**) C19, C25, C26 y C8 están cercanos, C2 un poco más distante, y C16 es el que más se aleja. Se encuentran más próximos a los testigos Nacional EET-103 y Trinitario EET-111.

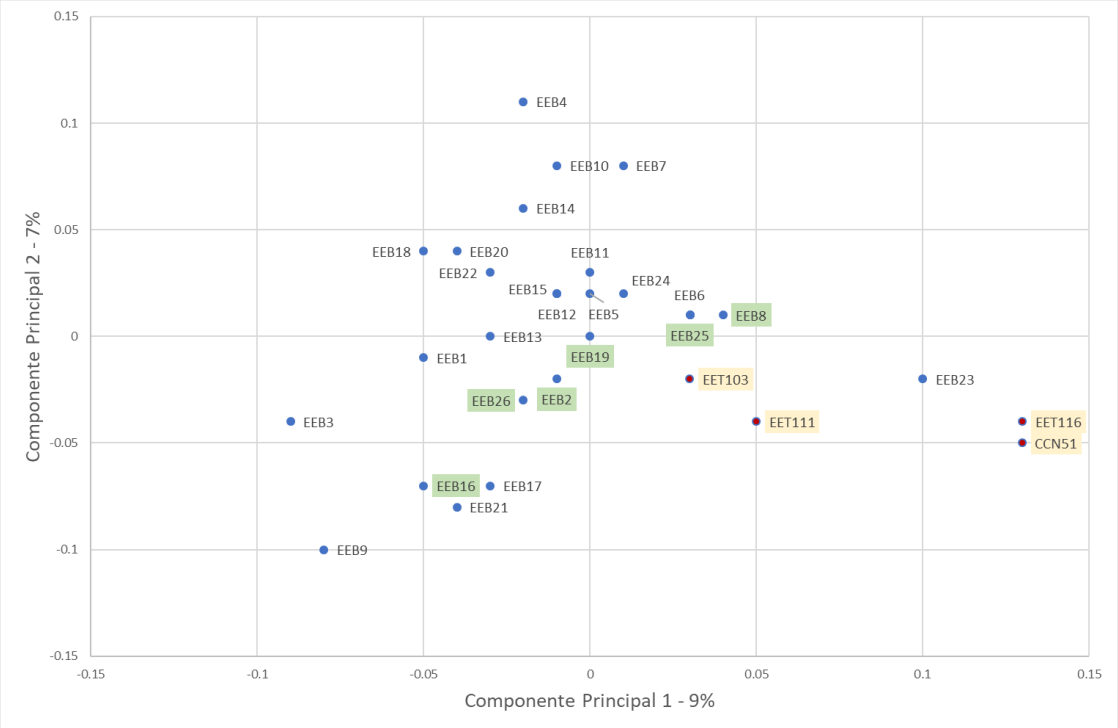

Gráfico 38-3: Componentes principales 1 y 2 del consenso morfológico-molecular. Explican el 16% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

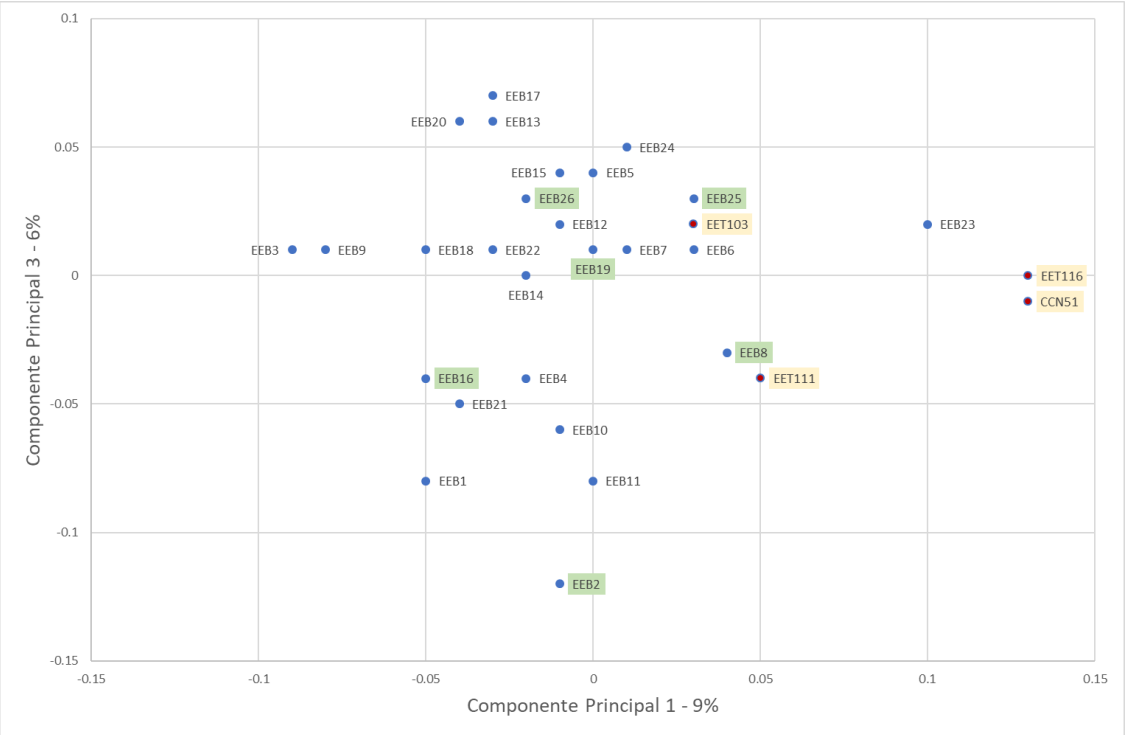

Gráfico 39-3: Componentes principales 1 y 3 del consenso morfológico-molecular. Explican el 15% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

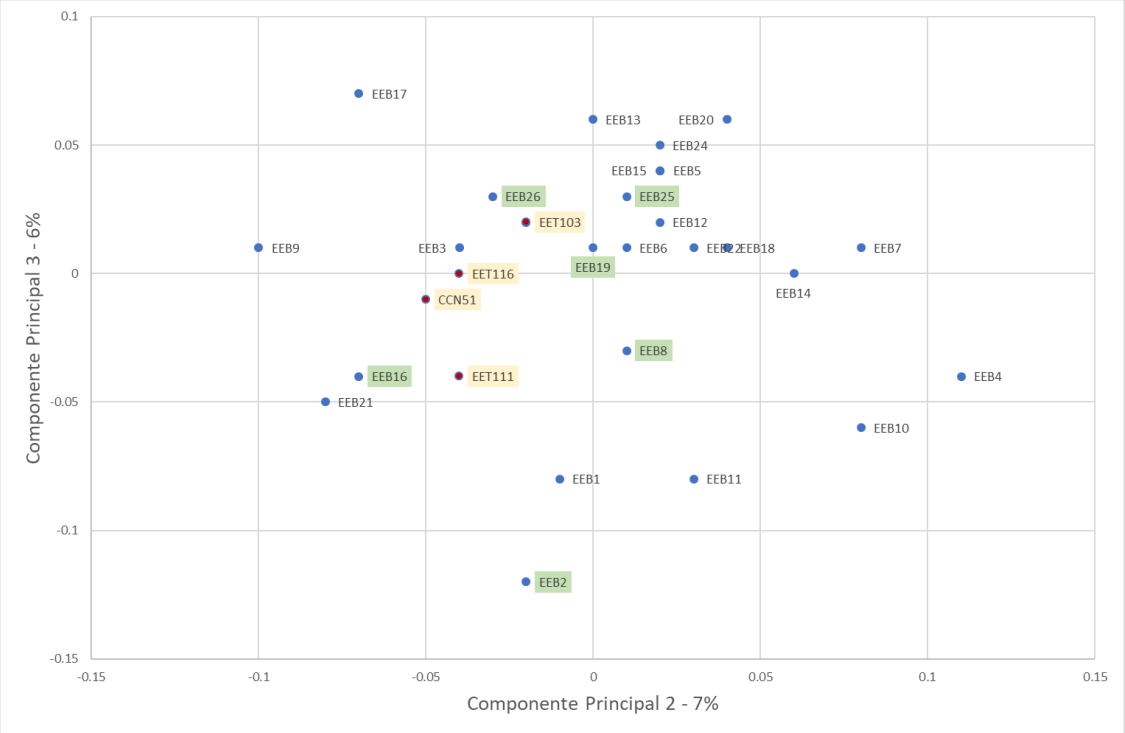

Gráfico 40-3: Componentes principales 2 y 3 del consenso morfológico-molecular. Explican el 13% de la variabilidad.

Realizado por: Gabriela J. Obregón O. 2018.

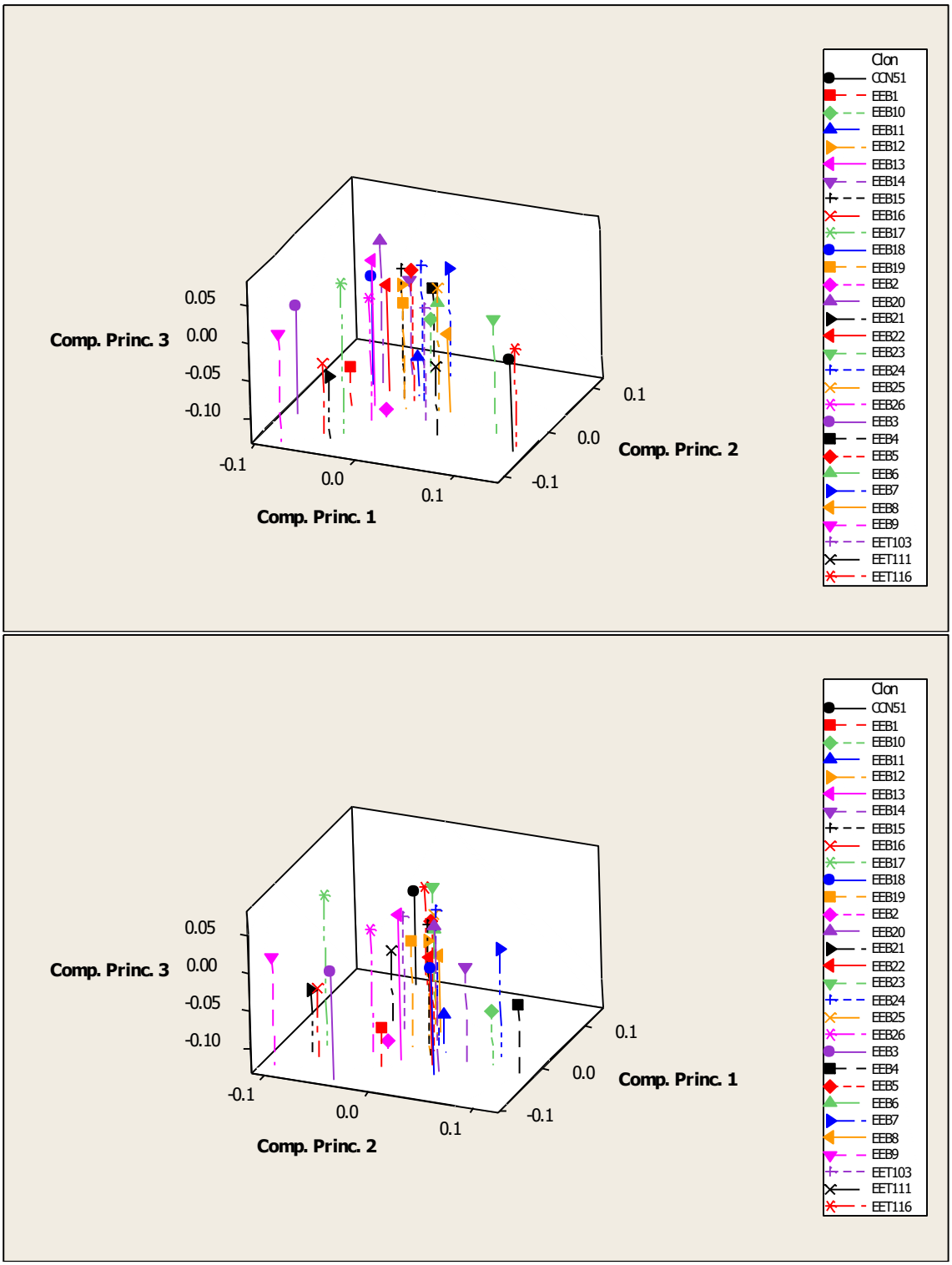

Gráfico 41-3: Visualización en 3D con los 3 primeros componentes principales del consenso morfológico-molecular, desde diferentes perspectivas.

Realizado por: Gabriela J. Obregón O. 2018.

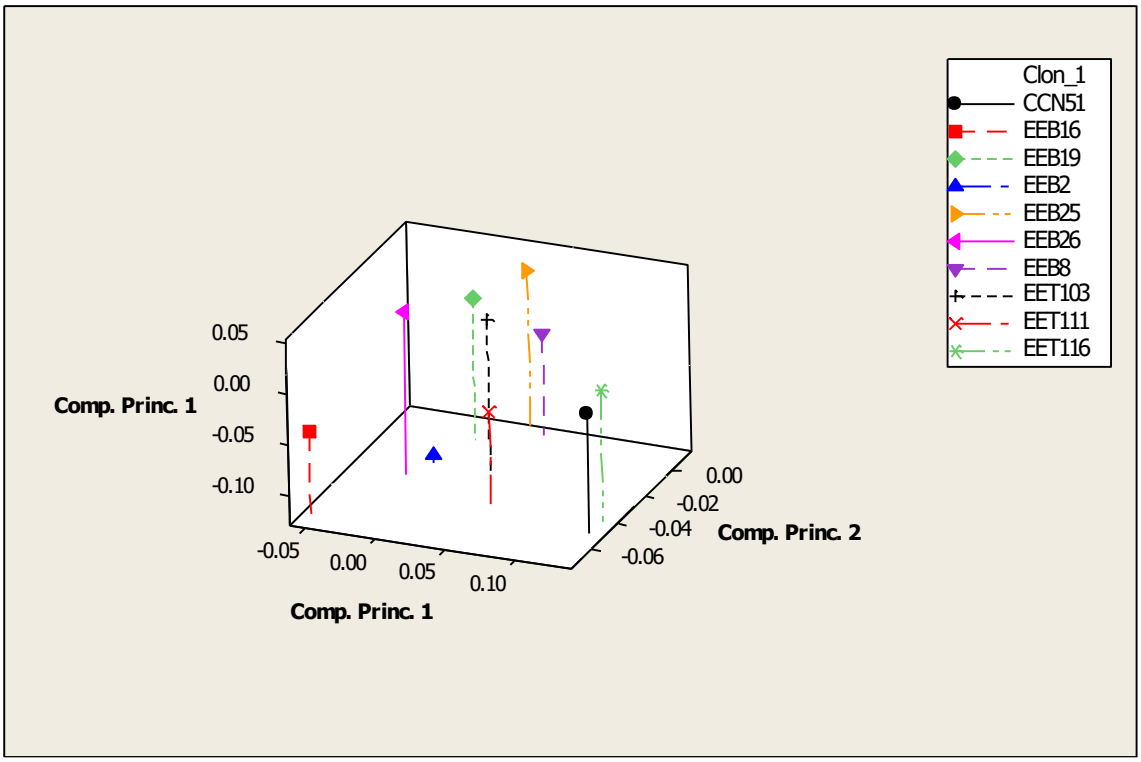

Gráfico 42-3: Gráfica en 3D con enfoque en los 6 de tipo Criollo de más alto rendimiento, en comparación con los testigos.

Realizado por: Gabriela J. Obregón O. 2018.

3.8 Post-proceso: Análisis de conglomerados

Se presentan los dendrogramas por agrupamiento jerárquico de Ward para cada configuración individual y para cada configuración consenso para determinar agrupaciones.

En el **Gráfico 43-3** se aprecian 2 grandes grupos diferenciados: C23, EET-116 y CCN-51 (dos de los testigos) en un grupo, y los demás en otro. Con los testigos restantes EET-111 (Trinitario) y EET-103 (Nacional) se agruparon C8, C7, C5, C24, C25 y C19. De los Criollos de alto rendimiento, C19 y C25 son los más similares. Los demás clones se pueden considerar más Criollos, especialmente los conglomerados 9, 6, 2 y 4.

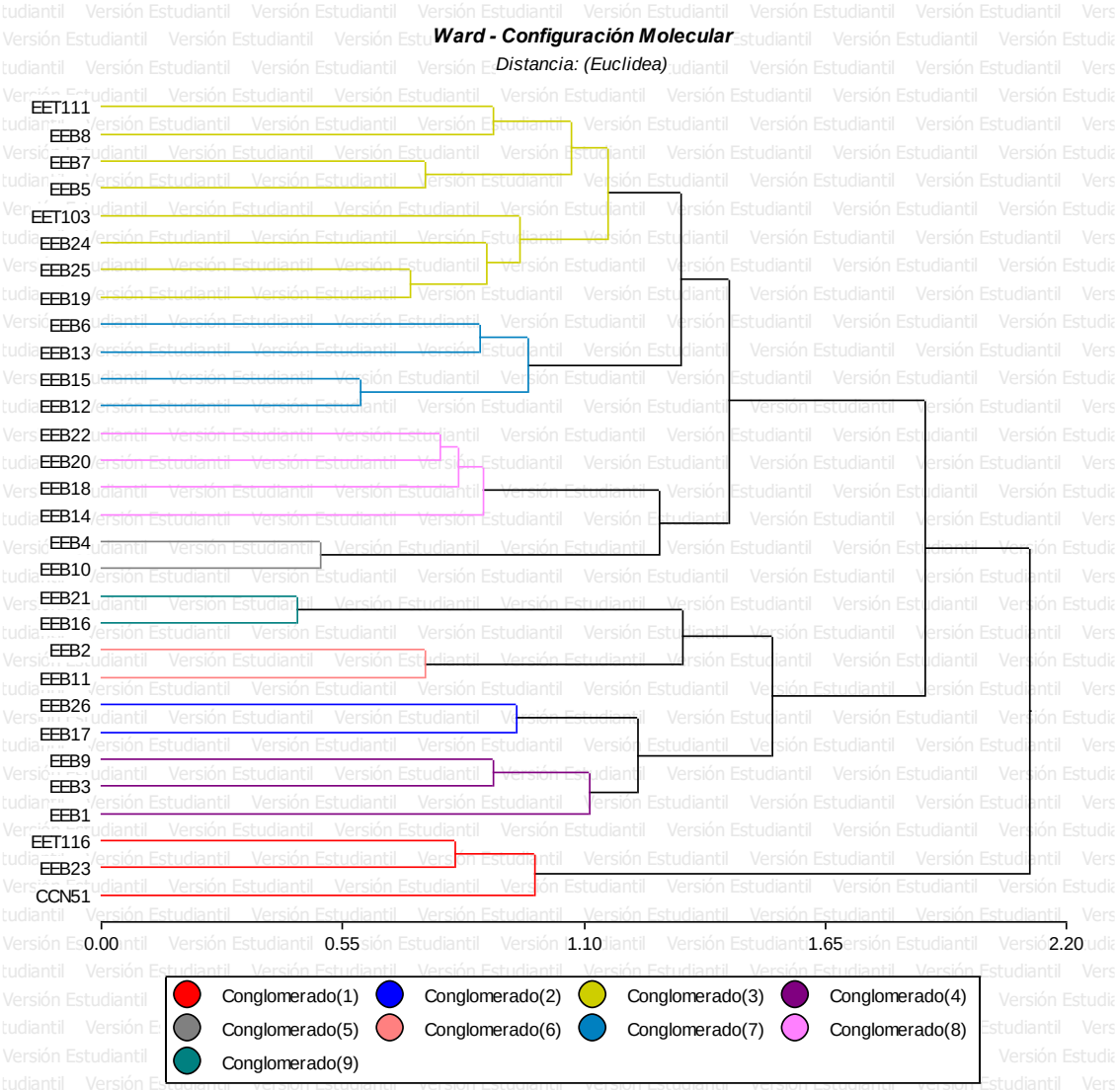

Gráfico 43-3: Dendrograma de la solución por coordenadas principales molecular.

Realizado por: Gabriela J. Obregón O. 2018.

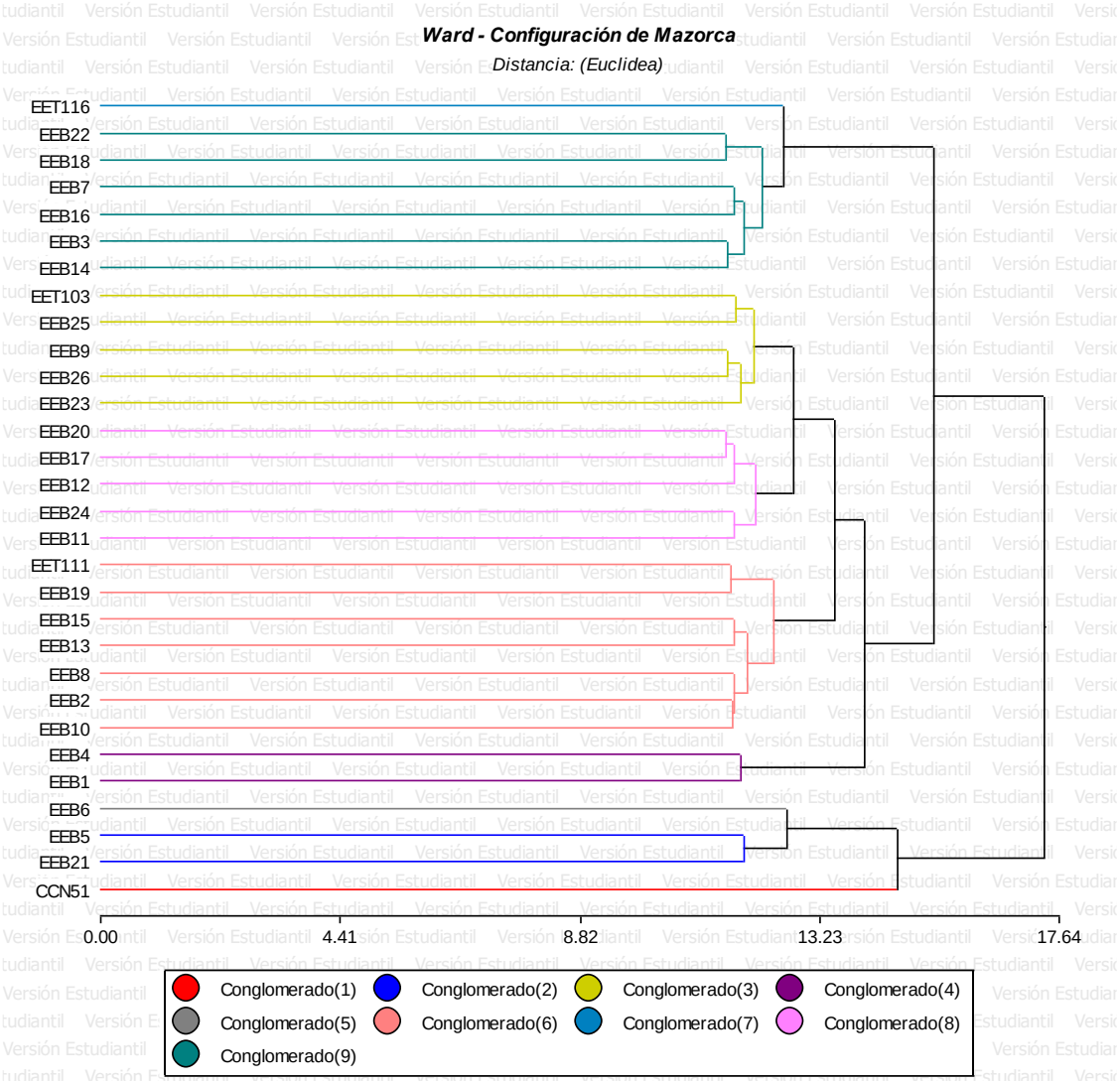

Gráfico 44-3: Dendrograma de la solución por coordenadas principales de mazorca.

Realizado por: Gabriela J. Obregón O. 2018.

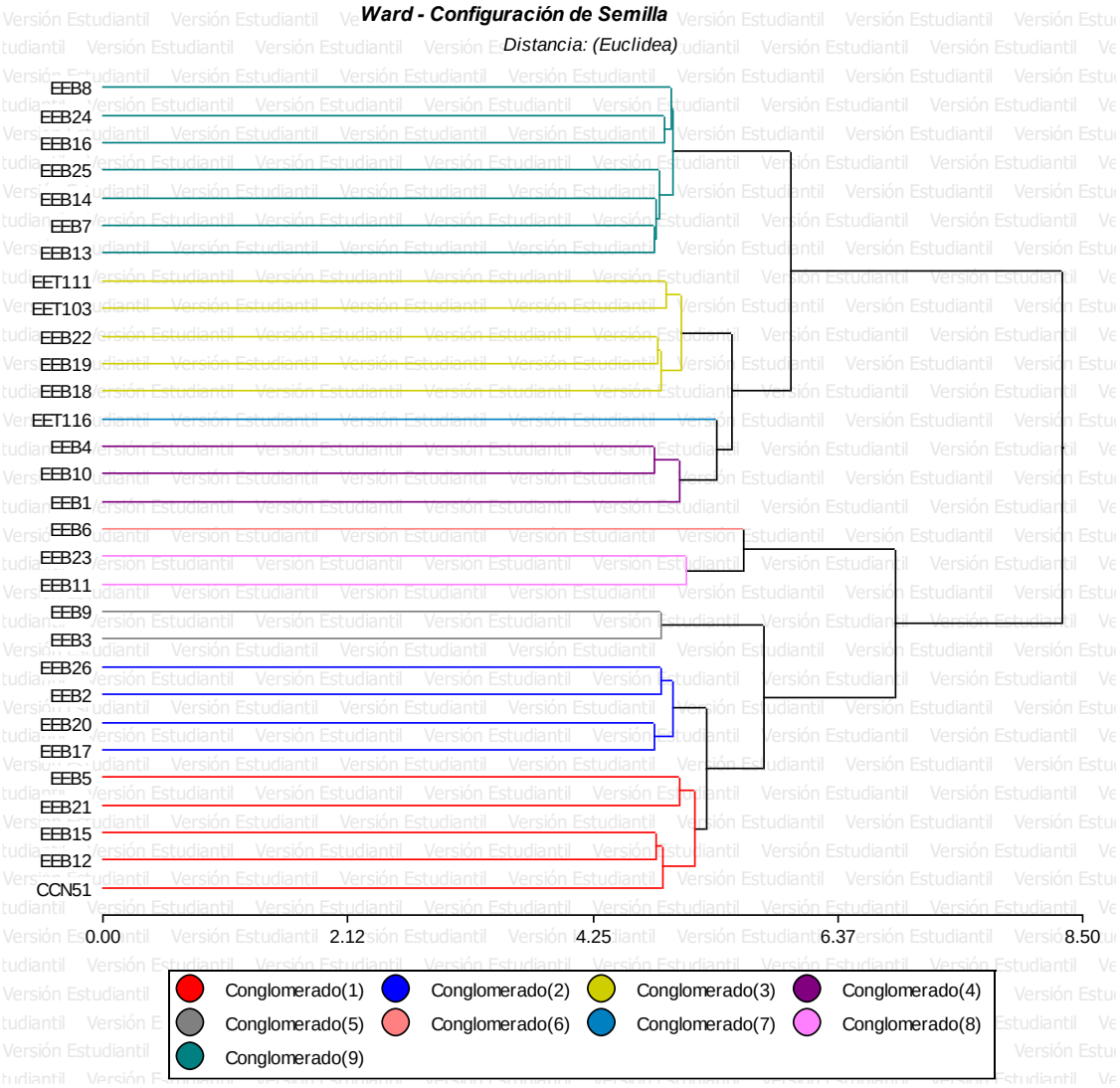

Gráfico 45-3: Dendrograma de la solución por coordenadas principales de semilla.

Realizado por: Gabriela J. Obregón O. 2018.

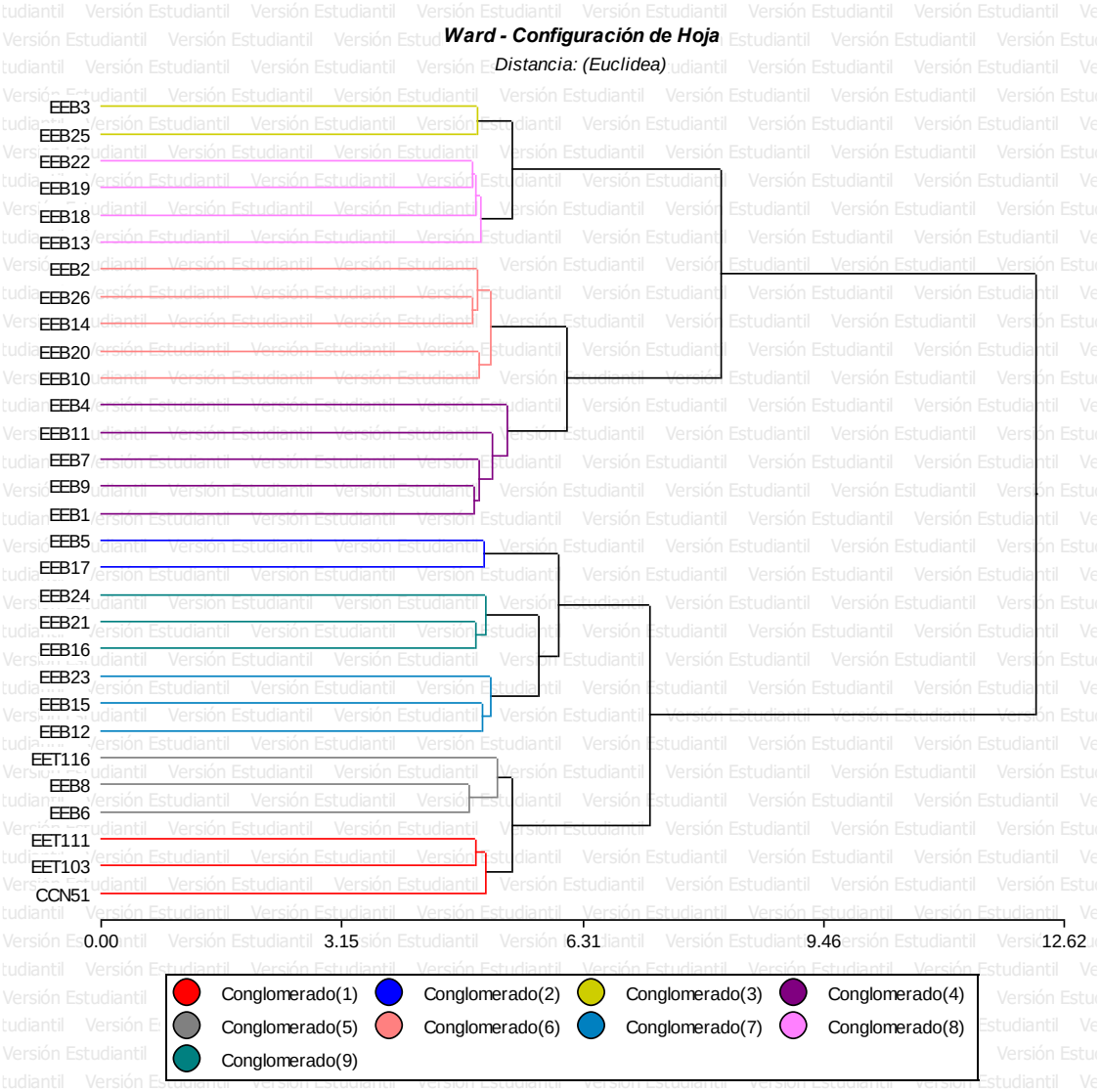

Gráfico 46-3: Dendrograma de la solución por coordenadas principales de hoja.

Realizado por: Gabriela J. Obregón O. 2018.

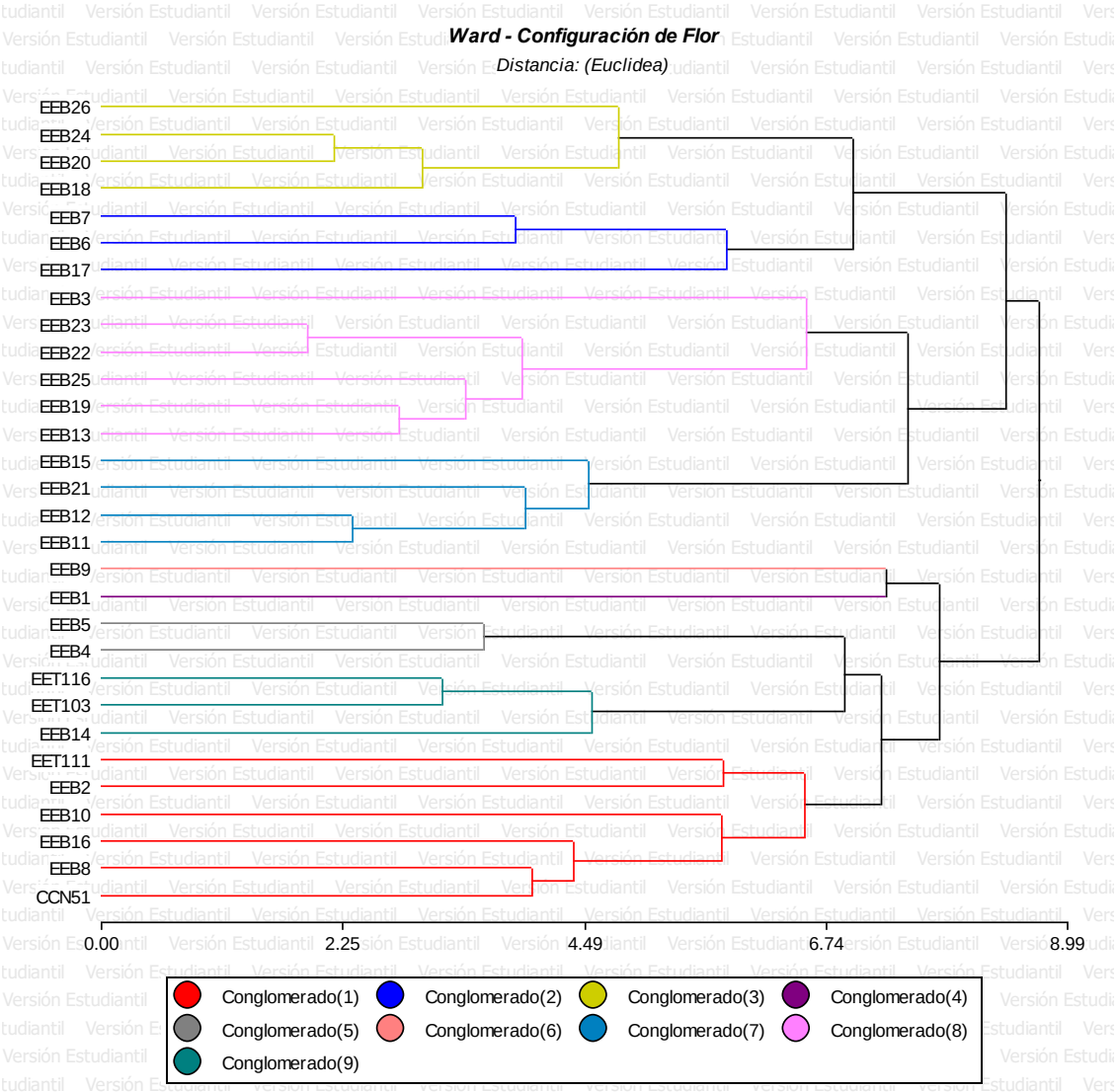

Gráfico 47-3: Dendrograma de la solución por coordenadas principales de flor.

Realizado por: Gabriela J. Obregón O. 2018.

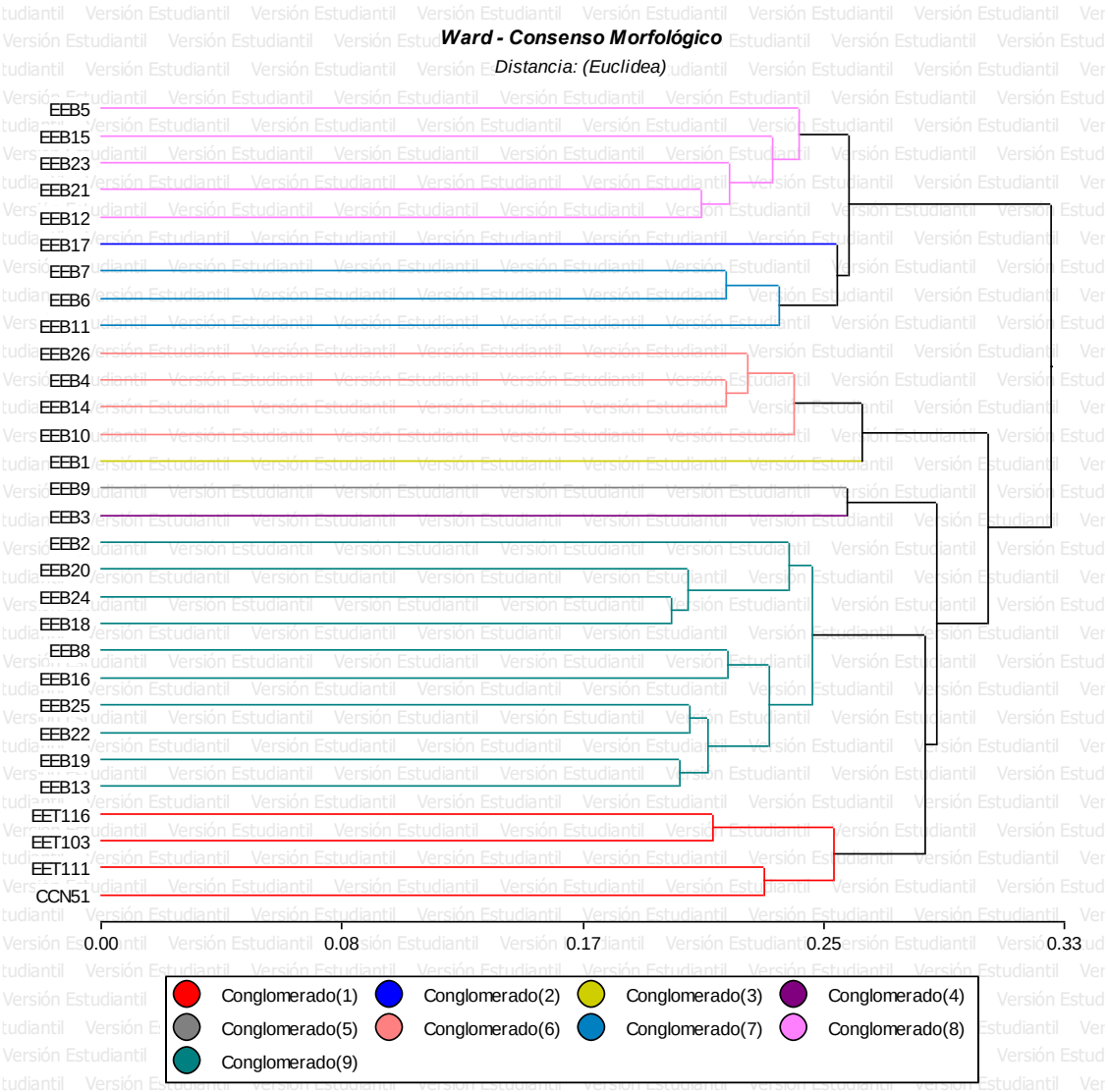

Gráfico 48-3: Dendrograma de la solución por componentes principales del consenso morfológico.

Realizado por: Gabriela J. Obregón O. 2018.

En el consenso morfológico los 4 clones testigo se diferencian claramente del resto, conformando el conglomerado 1. En el conglomerado más cercano, el 9, se encuentran C2, C8, C16, C25 y C19 (5 de los 6 de más alto rendimiento).

El conglomerado 1 (**Gráfico 49-3**), que contiene al Forastero C23, contiene también a CCN-51 y C23. El conglomerado 2, que contiene al Trinitario EET-111, contiene también a C26 y C17. El conglomerado 8, que contiene al Nacional EET-103, contiene también a C24, C20, C22, C18 y C14. Los demás clones se distancian más de los testigos.

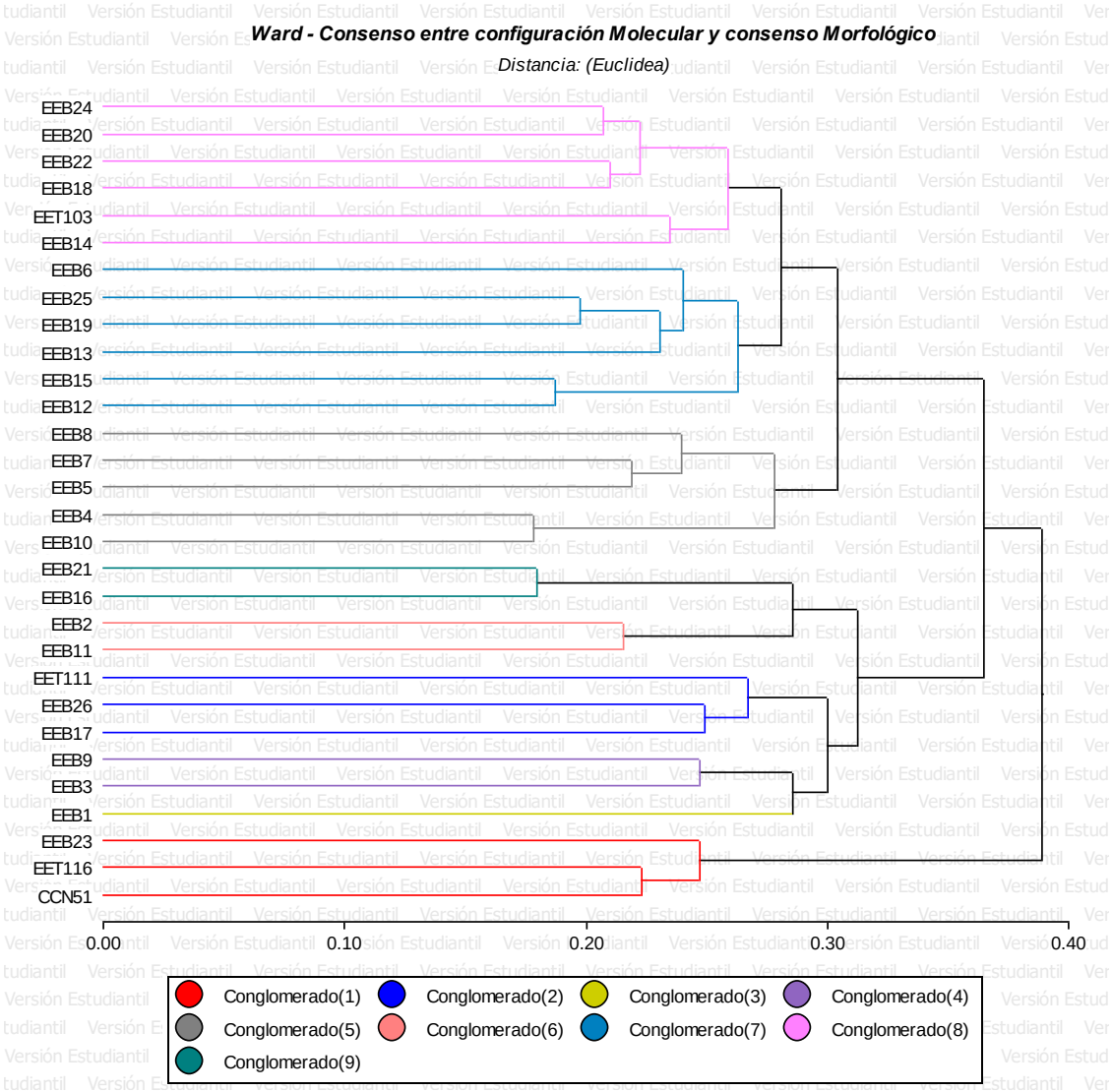

Gráfico 49-3: Dendrograma de la solución por componentes principales del consenso final.

Realizado por: Gabriela J. Obregón O. 2018.

3.9 Similitudes entre clones

Se presenta en forma de tabla los conglomerados que fueron determinados por los dendrogramas bajo cada uno de los enfoques considerados: marcadores moleculares, mazorca, semilla, hoja, flor, consenso morfológico y consenso morfológico-molecular. (Se muestran en color rojo los testigos y en color verde los de tipo Criollo que destacan en rendimiento).

Tabla 23-3: Conglomerados del dendrograma obtenido de la configuración molecular.

Marcadores moleculares SSR								
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
EET-111	C6	C22	C4	C21	C2	C26	C9	EET-116
C8	C13	C20	C10	C16	C11	C17	C3	C23
C7	C15	C18					C1	CCN-51
C5	C12	C14						
EET-103								
C24								
C25								
C19								

Realizado por: Gabriela J. Obregón O. 2018.

Tabla 24-3: Conglomerados del dendrograma obtenido de la configuración de mazorca.

Mazorca								
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
EET-116	C22	EET-103	C20	EET-111	C4	C6	C5	CCN-51
	C18	C25	C17	C19	C1		C21	
	C7	C9	C12	C15				
	C16	C26	C24	C13				
	C3	C23	C11	C8				
	C14			C2				
				C10				

Realizado por: Gabriela J. Obregón O. 2018.

Tabla 25-3: Conglomerados del dendrograma obtenido de la configuración de semilla.

Semilla								
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
C8	EET-111	EET-116	C4	C6	C23	C9	C26	C5
C24	EET-103		C10		C11	C3	C2	C21
C16	C22		C1				C20	C15
C25	C19						C17	C12
C14	C18							CCN-51
C7								
C13								

Realizado por: Gabriela J. Obregón O. 2018.

Tabla 26-3: Conglomerados del dendrograma obtenido de la configuración de hoja.

Hoja								
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
C3	C22	C2	C4	C5	C24	C23	EET-116	EET-111
C25	C19	C26	C11	C17	C21	C15	C8	EET-103
	C18	C14	C7		C16	C12	C6	CCN-51
	C13	C20	C9					
		C10	C1					

Realizado por: Gabriela J. Obregón O. 2018.

Tabla 27-3: Conglomerados del dendrograma obtenido de la configuración de flor.

Flor								
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
C26	C7	C3	C15	C9	C1	C5	EET-116	EET-111
C24	C6	C23	C21			C4	EET-103	C2
C20	C17	C22	C12				C14	C10
C18		C25	C11					C16
		C19						C8
		C13						CCN-51

Realizado por: Gabriela J. Obregón O. 2018.

Tabla 28-3: Conglomerados del dendrograma obtenido del consenso morfológico.

Consenso morfológico								
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
C5	C17	C7	C26	C1	C9	C3	C2	EET-116
C15		C6	C4				C20	EET-103
C23		C11	C14				C24	EET-111
C21			C10				C18	CCN-51
C12							C8	
							C16	
							C25	
							C22	
							C19	
							C13	

Realizado por: Gabriela J. Obregón O. 2018.

Tabla 29-3: Conglomerados del dendrograma obtenido del consenso entre configuración molecular y consenso morfológico.

Consenso morfológico - molecular								
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
C24	C6	C8	C21	C2	EET-111	C9	C1	C23
C20	C25	C7	C16	C11	C26	C3		EET-116
C22	C19	C5			C17			CCN-51
C18	C13	C4						
EET-103	C15	C10						
C14	C12							

Realizado por: Gabriela J. Obregón O. 2018.

3.10 Discusión

En lo que respecta al análisis de datos morfológicos, la metodología empleada en este estudio fue diferente a la desarrollada por el autor Pésantez (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014), motivo por el cual los resultados no coincidieron 100% a pesar de haber partido del mismo conjunto de datos morfológicos, como se muestra en el **Gráfico 50-3**, donde se presentan los dos dendrogramas obtenidos con ambos análisis. Se aprecia que, al hacer una comparación con respecto al dendrograma 1, sólo para algunos clones se

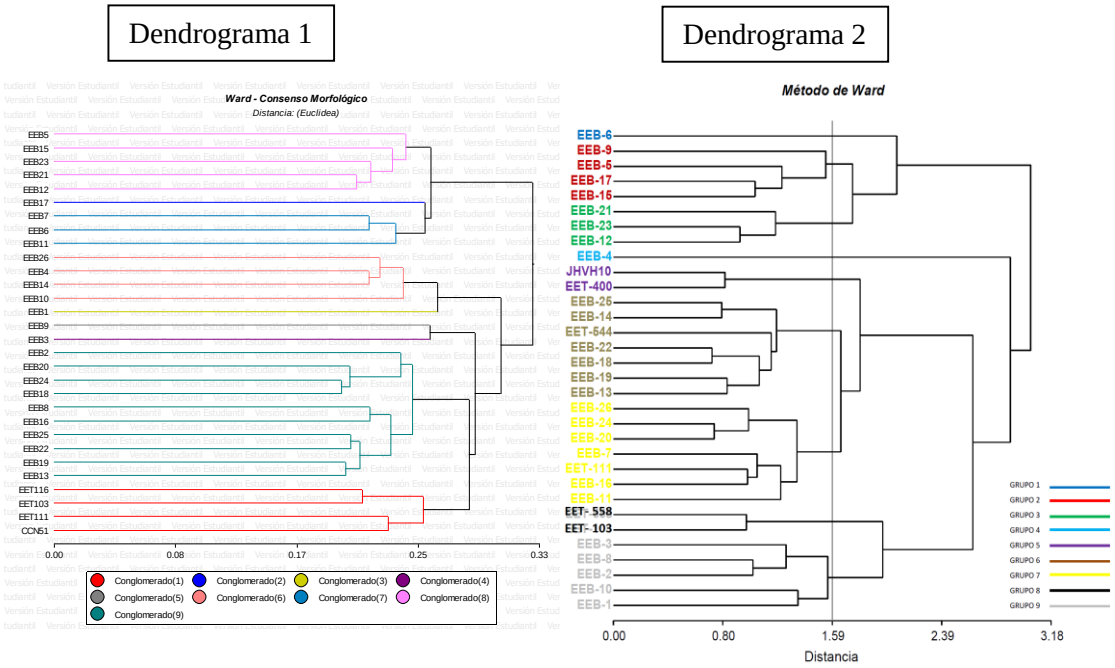

Gráfico 50-3: Derecha: resultado del presente análisis con testigos EET-103, EET-111, CCN-51 y EET-116. Izquierda: resultado de Pesántez (2014) con testigos EET-103, EET-111, JHVH-10, EET-400, EET-544, EET-558.

Realizado por: Gabriela J. Obregón O. 2018.

coincidió al asignarlos cercanos en un mismo conglomerado, así, los que coinciden y no en un mismo conglomerado son:

- Coinciden: C5, C15, C23, C21, C12, C17 y C6. No coinciden: C7 y C11.
- Coinciden: C26 y C14. No coinciden: C4[11], C10 y C1.
- No coinciden: C9 y C3.
- Coinciden: C20, C24, C18, C16, C25, C22, C19, C13. No coinciden: C2 y C8.
- No coinciden: EET-111 y EET-103 (testigos en común).

Dicho autor conformó una matriz de tamaño 32×38 (26 clones tipo Criollo y 6 testigos, con 31 variables cuantitativas y 7 cualitativas) con más variables que individuos, donde el dato observado de cada variable cuantitativa fue el promedio por clon, y donde 7 de las variables fueron cualitativas categóricas en las cuales se asignó arbitrariamente un número entre 1 y 3 a cada categoría posible, escogiendo a la moda como valor representativo de cada clon, y son: Xh5.1, Xf1, Xf2, Xf3, Xs4, y dos variables no consideradas en el presente estudio: Forma de mazorca y Forma de semilla. En dicha matriz se desarrolló un Análisis de Componentes Principales, a partir de las componentes principales se aplicó análisis de conglomerados con el método de Ward y el coeficiente de similaridad de Gower, y con una prueba Chi Cuadrado se determinaron las variables cualitativas más discriminantes que contribuyeron a explicar la diversidad. Además, se hizo un Análisis de Varianza para determinar si existe diferencia estadística significativa en características cuantitativas y se calculó el coeficiente de variación. Finalmente se hizo una comparación de medias utilizando la prueba de Duncan para cada variable cuantitativa y se determinaron las variables discriminantes, lo cual se hizo al final, una vez determinados los grupos en el dendrograma (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014).

En el presente trabajo se hizo el análisis aprovechando la totalidad de los datos y no únicamente los promedios de cada variable para representar a cada clon, con el objeto de no perder información sobre la variabilidad total. Si se considera importante basarse en el coeficiente de variación para tener una idea de la variabilidad de la especie, éste puede diferir si se utilizan todos los datos, y al hacerlo la información es más real al contar con muchos más datos. Si bien se llegó a la misma conclusión sobre las variables que más variabilidad genética determinaron, la magnitud de los coeficientes de variación fue diferente. Esto se muestra en la sección 3.1.

[11] Se detectó un error: al revisar los datos que fueron empleados para obtener el dendrograma 2, se encontró que hubo un error en los datos de mazorca para el clon C4 (EEB-4), pues en los promedios de ese clon para las variables de mazorca se encontraron valores que están fuera del campo de variación normal para una mazorca (por ejemplo: 31.84 cm como promedio de largo de mazorca), a lo cual se atribuye que en el dendrograma 2 se ubique a EEB-4 en un conglomerado aislado de los demás.

Para saber qué variables cuantitativas detectan diferencias significativas entre clones y así tener una idea de las variables que más pudieron influir de forma univariada en la formación de conglomerados del final, se realizaron pruebas de hipótesis en cada una de las 33 variables cuantitativas utilizando la totalidad de los datos originales, y debido a que la mayoría de variables no cumplieron con varios supuestos del ANOVA, se recurrió a pruebas no paramétricas para mayor confianza en la validez de los resultados, al no requerir éstas del cumplimiento de supuestos de distribución de probabilidad, además de ser más apropiadas para muestras pequeñas.

Tabla 30-3: Comparación de resultados de pruebas para detectar diferencia significativa en las variables.

Variables	**Diferencia detectada con los métodos**		
	ANOVA (Pesántez 2014)	**Kruskal Wallis**	**ANOVA basado en rango**
1. Xm1	NS	NS	***S***
2. Xm2	NS	***S***	***S***
3. Xm3	***S***	NS	***S***
4. Xm4	***S***	NS	***S***
5. Xm5	***S***	***S***	***S***
6. Xm6	***S***	***S***	***S***
7. Xm7	***S***	***S***	***S***
8. Xm8	***S***	NS	***S***
9. Xm9	***S***	***S***	***S***
10. Xm10	***S***	NS	***S***
11. Xm11	NS	***S***	***S***
12. Xm12	***S***	***S***	***S***
13. Xm13	No disponible	NS	***S***
14. Xs1	***S***	***S***	***S***
15. Xs2	***S***	***S***	***S***
16. Xs3	***S***	***S***	***S***
17. Xh1	NS	***S***	***S***
18. Xh2	***S***	***S***	***S***
19. Xh3	***S***	NS	***S***
20. Xh4	NS	***S***	***S***
21. Xh5.1	No disponible	NS	***S***
22. Xh6	NS	***S***	***S***
23. Xh7	NS	***S***	***S***
24. Xf4	NS	***S***	***S***
25. Xf5	NS	***S***	***S***
26. Xf6	NS	***S***	***S***
27. Xf7	NS	***S***	***S***
28. Xf8	***S***	***S***	***S***
29. Xf9	NS	***S***	***S***

30. Xf10	NS	***S***	No disponible
31. Xf11	NS	***S***	***S***
32. Xf12	NS	***S***	***S***
33. Xf13	***S***	***S***	***S***

NS = No Significativa

S = Significativa

Realizado por: Gabriela J. Obregón O. 2018.

Se aplicó la prueba de Kruskal Wallis y la prueba de Análisis de Varianza de Una Vía Basada en Rango (robusto), las cuales detectaron diferencia significativa. Por su parte, Pezántez (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014) determinó 15 variables discriminantes en las cuales comparó las medias de los grupos que se formaron en el análisis de conglomerados final. Se hace una comparación en la **Tabla 30-3**.

De forma general se corrobora que todas las variables utilizadas detectan diferencias entre los clones, especialmente con el ANOVA basado en Rango. Llama la atención que con las pruebas no paramétricas y utilizando todos los datos observados de flores de forma univariada, se encuentra que casi todas las variables cuantitativas de flor detectan al menos un par de clones que difieren, en contraste con el análisis de Pesántez (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014) que únicamente determina 2 variables cuantitativas de flor como discriminantes. Puede ser un indicio de que las variables de flor son importantes para discriminar clones, pero cuando se utilizan todos los datos. A pesar de esto, por motivos de carecer de una matriz con todas las variables observadas sobre todas las flores, en este trabajo tampoco se pudo utilizar la totalidad de datos de flores para análisis multivariante, por lo que se utilizaron únicamente los promedios por clon.

Para análisis multivariante morfológico, al ser el clon la unidad de análisis de interés, se encontró una forma de analizar no solo los promedios por clon sino la totalidad de los datos observados mediante el cálculo de la distancia de Mahalanobis entre pares de poblaciones, lo que permitió enfocarse en cada clon como unidad de estudio. Otra diferencia estuvo en que no se analizaron todas las variables morfológicas en una misma matriz, sino que se destinó una matriz por cada unidad de análisis (mazorca, semilla, hoja y flor), de las que se obtuvo una matriz de distancias entre clones para obtener coordenadas principales, así se pudo observar cómo se agrupan los clones de acuerdo a cada unidad de análisis por separado. Con APG se obtuvo un consenso morfológico entre las coordenadas principales de mazorca, semilla, hoja y flor que fue alto: 83.7%. La matriz consenso entregada por el programa Infostat fue en forma de componentes principales. A esta matriz resultante de este método, a la que se aplicó análisis de conglomerados, en los grupos que se determinaron con agrupamiento jerárquico de Ward empleando la distancia

euclídea (al considerar que es más apropiada luego de las transformaciones a las que se sometieron los datos), los clones testigo (no Criollos) y que tienen alto rendimiento, se diferenciaron notablemente de los clones de tipo Criollo, y de forma general se ubicaron más cercanos a los de tipo Criollo de más alto rendimiento y más lejanos a los de menor rendimiento, lo cual tiene sentido y fue validado por el director del Programa de Cacao (se adjunta el audio en la *nube* (link: https://onedrive.live.com/?authkey=%21AIJE2OcZQnWrOgE&id=CA77A5953C07ED0C%21140440&cid=CA77A5953C07ED0C)). De esto se podría deducir que en este análisis las características morfológicas (especialmente las de mazorca, semilla y hojas que contaron con un gran número de datos) pareciera que fueron más apropiadas para formar conglomerados por rendimiento, al estar directamente asociadas a éste, antes que separar los clones por grupo genético, lo cual sí lo hace de forma determinante el análisis molecular.

En este estudio se aportó con información complementaria al aspecto morfológico: datos proporcionados por 20 marcadores moleculares SSR, lo que permitió agrupar los clones con respecto a su ADN y grupo genético, y finalmente llegar a un consenso final entre el aspecto morfológico y molecular para obtener una clasificación más completa y eficaz en la cual los fitomejoradores puedan basarse para tomar decisiones acerca de los cruces a realizar, o para corroborar las decisiones ya tomadas.

Con esto, finalmente se realizó un consenso entre el aspecto morfológico y molecular, el cual fue de 94.3%. Un estudio donde se analizaron características moleculares (con SSR) y morfológicas en cacao nacional de Bolivia con APG fue el de (JULY MARTINEZ 2007) donde se obtuvo un consenso de 70.8%, el cual fue más bajo que el del presente trabajo, quizá debido a una mayor diversidad del material de estudio, pero aun así se considera alto.

CONCLUSIONES

- La unidad de análisis más variable fue la mazorca, con coeficientes de variación entre 13% y 50%.
- A nivel molecular los clones de tipo Criollo se asemejaron entre ellos, y en general se asemejaron más a los testigos Nacional y Trinitario. Se distanciaron del clon Forastero, de su descendiente directo CCN-51, así como del único clon de tipo Criollo que se apartó: C23. Con respecto a los testigos, estuvieron en un mismo conglomerado EET-111 - C8 - C7 - C5, y EET-103 - C24 - C25 - C19.
- En el consenso morfológico obtenido, las variables de las unidades de análisis morfológicas que contaron con una gran cantidad de datos, como mazorca, semilla y hoja, parecen haber agrupado los clones más en función de su rendimiento que de su grupo genético, lo cual se refleja, por ejemplo, en que EET-103 (Nacional) o EET-111 (Trinitario) se asignó a un mismo conglomerado con los demás testigos, cuando cada uno representa a un diferente grupo genético, pero tienen en común la característica de tener un alto rendimiento.
- Para CCN-51, que es ampliamente cultivado por su alta productividad, fue notorio su distanciamiento en cuanto a variables de mazorca, destacando también en variables de hoja. No fue así para variables de semilla ni de flor. Además, molecularmente en las coordenadas principales se ubicó entre sus padres EET-111 y EET-116.
- En el consenso morfológico se distanciaron del grupo central los clones C6, C5, C9, C3, C4, C1 y CCN-51. En conglomerados, los testigos se diferenciaron claramente de los 26 de tipo Criollo, que conformaron un grupo aparte. Los resultados obtenidos con Agrupamiento jerárquico de Ward a partir del consenso del Análisis Procrustes Generalizado no fueron completamente similares a los obtenidos por Pesántez (PESANTEZ REYES, Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas 2014) con agrupamiento jerárquico de Ward y coeficiente de similitud de Gower, pues el tratamiento que se dio a los datos fue diferente: en este trabajo se emplearon todos los datos recolectados y no únicamente los promedios por clon (excepto para flores), se empleó la distancia de Mahalanobis entre poblaciones para considerar las correlaciones lineales entre las variables, y para el agrupamiento final se partió de un consenso (morfológico) entre mazorca, semilla, hoja y flor.
- En el consenso morfológico-molecular:
    - Permanecieron separados en un grupo el clon Forastero EET-116 - CCN-51 - C23 (lo que se atribuye al peso que tuvo la caracterización molecular que los distanciaba notoriamente).

    - Los más cercanos al testigo Trinitario EET-111 fueron C26 y C17.
    - Los más cercanos al testigo Nacional EET-103 fueron: C14, C18, C22, C20 y C24.
    - Formaron grupos pequeños C16 - C21, y C4 - C10, lo que también se podría atribuir al peso que tuvo la alta similitud de éstos en la caracterización molecular.
- Tanto en la configuración molecular, como en la de semilla, en el consenso morfológico, y en el consenso morfológico molecular, los clones C3 y C9 conformaron un conglomerado, lo cual llama la atención al comprobar en el campo que en dichos materiales hay presencia de semillas blancas.
- La agrupación morfológica no fue igual a la agrupación molecular. En el consenso final se obtuvo una agrupación más efectiva de los clones al separarlos según marcadores moleculares y morfológicos simultáneamente. El porcentaje de consenso morfológico-molecular fue alto: 94.3%.
- Respecto al seguimiento a los 6 tipo Criollo de mayor rendimiento, en la caracterización molecular los más similares fueron C19 y C25. En el consenso morfológico se ubicaron en un mismo conglomerado C2 - C8 - C16 - C19 - C25, y C26 se ubicó en otro. Mientras que en el consenso morfológico-molecular C26 estuvo en un mismo grupo con EET-111; C19 y C25 pertenecieron a otro conglomerado; y C8, C16 y C2 se ubicaron cada uno en un conglomerado diferente. Ninguno de ellos se agrupó con el Nacional EET-103.
- Las pruebas no paramétricas confirmaron que las variables morfológicas aquí consideradas son apropiadas para diferenciar entre los clones.
- Se considera que la técnica de Análisis Procrustes Generalizado aplicada para tratar configuraciones de diferente naturaleza (morfológica y molecular) fue apropiada para este caso y dio buenos resultados, pues los porcentajes de consenso fueron altos, y a partir de sus resultados se pudo agrupar de forma más efectiva los clones de cacao con respecto a las características estudiadas. Esto se ve reflejado en el hecho de que se evidenció una separación en cuanto a grupos genéticos, considerando simultáneamente el aspecto morfológico, pues ambos tipos de información son de interés e importancia agronómica.

RECOMENDACIONES

- Los 6 clones de mayor rendimiento, en general son semejantes, no se alejan mucho entre ellos tanto en las caracterizaciones individuales como en los consensos (el que más se distancia es C16, que tiene buenas características organolépticas), por lo que se recomienda decidir cuál de ellos utilizar para el mejoramiento genético, considerando otros aspectos que mejor convengan según el criterio del fitomejorador, como resistencia a enfermedades o características de calidad, para generar nuevas plantas mejoradas, resistentes y con características de fino y de aroma.
- Para futuros estudios similares de caracterización morfológica y molecular, se recomienda analizar la totalidad de los datos disponibles de variables morfológicas para reflejar de mejor manera la variabilidad total y aprovechar toda la información, en donde se puede emplear la distancia de Mahalanobis entre poblaciones para finalmente enlazar las poblaciones o clones con respecto a sus variables morfológicas y moleculares.
- Investigar sobre el índice de similitud más adecuado según la naturaleza de las variables moleculares y de la especie en estudio, para que los resultados tengan sentido para fines genéticos. En este caso se consideró apropiado el recomendado por Kosman y Leonard.
- Para estudiar características morfológicas de flores, se recomienda medir todas las variables sobre una misma flor para obtener una matriz de p variables medidas sobre n flores y poder realizar análisis multivariante con mayor información para cada clon. Además, procurar que, sea cual sea la matriz, se cumpla que $n \geq 10p$ o que $n \geq 20p$, de ser posible, para que sea más apropiado analizar con técnicas multivariantes.
- Al caracterizar mazorcas, tratar de hacerlo en lo posible al mismo tiempo de haber sido cosechadas para evitar introducir errores relacionados a la pérdida de humedad, que afecta a algunas variables.
- Aplicar Análisis Procrustes Generalizado no solo para caracterizar otros cultivos con conjuntos de variables de diferente naturaleza, en este caso morfológicas y moleculares medidas sobre un conjunto de plantas, sino también en cualquier situación en la que se disponga de 2 o más conjuntos de variables de diferente tipo, medidas sobre el mismo conjunto de individuos, ya que esta técnica permite consensuar las diferencias entre los tipos de variables y determinar una mejor clasificación al considerar varios aspectos simultáneamente, o bien ver qué tan alto es el consenso que puede haber entre los tipos de variables así como entre los individuos. Algunos de los campos de aplicación donde se la aplica usualmente son: análisis de perfil sensorial, geometría morfométrica, análisis de forma y análisis de imágenes.
- También se puede hacer este análisis con el paquete FactoMineR, que, en caso de necesitarlos, entrega más resultados para APG que Infostat al proporcionar no sólo las

tablas de PANOVA por configuración y por individuo, sino también coeficientes e índices de similaridad entre configuraciones parciales, las configuraciones parciales antes y después de aplicar la transformación Procrustes, correlaciones entre configuraciones parciales iniciales y dimensiones de consenso, y PANOVA por dimensión, y es más flexible porque permite establecer un número deseado de iteraciones, así como de límite de tolerancia para las mismas. A pesar de esto último, los resultados en común fueron prácticamente los mismos que los entregados por Infostat en cuanto a porcentaje de consenso.

- Hacer un estudio a futuro que permita determinar con qué grado de confiabilidad se pueden generalizar los resultados del análisis de conglomerados, lo cual quedó como interrogante.

BIBLIOGRAFÍA

Asociación Nacional de Exportadores de Cacao - Ecuador (ANECACAO). *El cacao, uno de los más significativos símbolos de nuestro país.* [en línea] (consulta: 26 de 04 de 2018) Disponible en: http://www.anecacao.com/index.php/es/quienes-somos/cacao-en-ecuador.html.

Asociación Nacional de Exportadores de Cacao - Ecuador (ANECACAO). *Estadísticas de exportación.* [en línea] (Consulta: 26 de 04 de 2018). Disponible en: http://www.anecacao.com/index.php/es/estadisticas/estadisticas-actuales.html (último acceso: 26 de 04 de 2018).

Azofeifa Delgado, Alvaro. «Uso de marcadores moleculares en plantas: Aplicaciones en frutales del trópico.» *Agronomía Mesoamericana, vol.* 2, nº 17 (2006) pp: 221-242.

Bekele, F.; Butler D. R. «Proposed list of cocoa descriptors for characterisation.» *Proceedings of the CFC/ICCO/IPGRI Project Workshop* (IPGRI), 2000: 41-48.

Bramardi, Sergio Jorge; et. al. «Simultaneous agronomic and molecular characterization of genotypes via the Generalised Procrustes Analysis: an application to cucumber.» *Crop Science* 45, nº 4 (2005): 1603–1609.

Bruno, Cecilia; Balzarini, Mónica. «Ordenaciones de material genético a partir de información multidimensional.» *Revista de la Facultad de Ciencias Agrarias* 42, nº 2 (2010): 183-200.

Calderón Cisneros, Juan Tarquino. *Mortalidad causada por tumores malignos en Ecuador, período 2005-2014: un estudio demográfico utilizando Análisis de Correspondencias Múltiple (ACM) y Análisis Procrustes Generalizado (AGP).* (tesis) (Maestría) Universidad de Salamanca - Departamento de Estadística, Salamanca, España. 2016.

Casaluker. *Cacao fino de aroma.* http://www.cacaofinodearoma.com/es/cacao-fino-de-aroma/ (último acceso: 26 de 04 de 2018).

Centro Agronómico y Tropical de Investigación y Enseñanza (CATIE). *International cacao cultivar catalogue.* Editado por JORGE SORIA V., & GUSTAVO A. ENRÍQUEZ. Vol. 6. Turrialba: CATIE. Perennial plant program, 1981.

Costa Tartara, Sabrina; et al. «Análisis simultáneo de variables morfológicas cuantitativas y marcadores moleculares para la caracteriación de accesiones nativas de quinoa del noroeste argentino.» *XVI Reunión Científica del Grupo Argentino de Biometría*, octubre 2011: 1-10.

Cuadras, Carles. «Distancias Estadísticas.» *Estadística Española* 30, nº 119 (1989): 295-378.

Cuadras, Carles. *Nuevos Métodos de Análisis Multivariante.* Barcelona: CMC Editions, 2012.

Demey, Jhonny R.; et al. «Medidas de distancia y de similitud.» *Valoración y análisis de la diversidad funcional y su relación con los servicios ecosistémicos*, s.f.: 47-59.

Demey, Jhonny R.; et al. «Using generalized procrustes analysis (GPA) to study the relationships between biochemical, molecular and morphological characterization in cassava collection.» *Conference: II Meeting Caribbean and Central American Region of the International Biometrics Society*, enero 2003.

Douglas Steel, Robert George; Torrie, James Hiram. *Principles and Procedures of Statistics.* New York: McGraw-Hill, 1960.

El Ciudadano. *Cacao fino de aroma codiciado por los grandes chocolateros del mundo.* 08 de abril de 2015. http://www.elciudadano.gob.ec/cacao-fino-de-aroma-codiciado-por-los-grandes-chocolateros-del-mundo/ (último acceso: 26 de 04 de 2018).

Gower, John C. «Generalized Procrustes Analysis.» *Psychometrika, vol.* 40, nº 1 (1975): 33-50.

Gower, John C.; Dijksterhuis, Garmt B. *Procrustes Problems.* Vol. Oxford Statistical Science Series 30. 30 vols. New York, United States: Oxford University Press Inc., 2004.

Grandon, Nancy G; et al. «Genetic diversity among alfalfa genotypes (Medicago sativa L.) of non-dormant cultivars using SSR markers and agronomic traits.» *Revista de la Facultad de Ciencias Agrarias* (Universidad Nacional de Cuyo) 45, nº 2 (2013): 181-195.

Guacho Abarca, Edison Fernando. *Caracterización agro-morfológica del maíz (Zea mays L.) de la localidad San José de Chazo.* Riobamba, Ecuador. (tesis) (Pre-grado) Escuela Superior Politécnica de Chimborazo - Facultad de Recursos Naturales - Escuela de Ingeniería Agronómica, Riobamba, Ecuador. 2014.

Hernandez Sampieri, Roberto; et al. *Metodología de la Investigación - Quinta Edición.* México D.F.: MGraw-Hill, 2010.

Hettmansperger, T. P.; Mckean, J. W. *Robust Nonparametric Statistical Methods, 2nd ed.* New York: Chapman-Hall, 2011.

Instituto Nacional Autónomo de Investigaciones Agropecuarias. «Informe técnico bianual 2009-2010 del Programa Nacional de Cacao y café de la Estación Experimental Tropical Pichilingue.» INIAP, Quevedo, Los Ríos, Ecuador, 2011, 134.

International Plant Genetic Resources Institute (IPGRI). *Análisis estadístico de datos de caracterización morfológica de recursos fitogenéticos. Boletín técnico no. 8.* Editado por TITO L., Franco; & HIDALGO, Rigoberto. Cali, Colombia: Instituto Internacional de Recursos Fitogenéticos (IPGRI), 2003.

Jana, Constanza; et al. «Morphological and genetic characterization among wild populations of copao (Eulychnia acida Phil.), cactus endemic to Chile.» *Chilean Journal of Agricultural Research* 77, nº 1 (Enero - Marzo 2017).

July Martinez, Windson. *Caracterización morfológica y molecular del Cacao Nacional Boliviano y de selecciones élites del Alto Beni, Bolivia.* (tesis) (Maestría) CATIE, Turrialba, Costa Rica. 2007.

Koebner, R. M.; Centre, John Innes. «Molecular Markers.» *Crop Improvement* (Elsevier), 2003: 140-146.

Kosman, E.; Leonard, K J. «Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species.» *Molecular Ecology* 14 (2005): 415-424.

Loor Solorzano, Rey Gastón. *Caracterización morfológica y molecular de 37 clones de cacao (Theobroma cacao L.) Nacional de Ecuador.* (tesis) (Maestría) Colegio de Postgraduados, Montecillo, Texcoco, Edo de México. 2002.

Lozada Vargas, Paúl Francisco. *Caracterización molecular de 42 accesiones de la Colección de Genotipos de Cacao Nacional (Theobroma cacao L.) de la EET Pichilingue, INIAP, mediante el uso de marcadores microsatélites (SSRs).* (tesis) (Pre-grado) Universidad de las Fuerzas Armadas (ESPE), Sangolquí, Ecuador. 2014.

Mahuad, Sabina L.; et al. «Preservation of Solanum pimpinellifolium genomic fragments in recombinant genotypes improved the fruit quality of tomato.» *Journal of Genetics* 92, nº 2 (agosto 2013): 195-203.

Ministerio De Agricultura Y Ganadería (MAGAP). *MAGAP impulsa proyecto de reactivación del Cacao Fino y de Aroma.* 2013. http://www.agricultura.gob.ec/magap-impulsa-proyecto-de-reactivacion-del-cacao-fino-y-de-aroma/ (último acceso: 26 de 04 de 2018).

Morillo Velastegui, Eduardo; Miño Castro, Gabriela. *"Marcadores moleculares en biotecnología agrícola: manual de procedimientos y técnicas en INIAP".* Vol. Manual No. 91. Quito, Ecuador: Instituto Nacional Autónomo de Investigaciones Agropecuarias, Estación Experimental Santa Catalina, 2011.

Peña, Daniel. *Análisis de datos multivariantes.* 2002.

PESANTEZ REYES, Adrean William. *Caracterización morfológica y de rendimiento de 26 clones de cacao (Theobroma cacao L.) considerando características de 6 genotipos identificados en la zona de Yaguachi provincia del Guayas.* (tesis) (Pre-grado) Universidad Agraria del Ecuador, Milagro - Ecuador. 2014.

Quiroz V., James, Soria V., Jorge. «Caracterización fenotípica del cacao nacional de Ecuador.» *Boletín técnico N. 41 - Estación Experimental Tropical Pichilingue* (INIAP), 1994: 1-16.

Quiroz Vera, James Gonzalo. *Caracterización molecular y morfológica de genotipos superiores de cacao nacional (Theobroma cacao L.) de Ecuador.* (tesis) (Maestría) Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), Turrialba, Costa Rica. 2002.

Quiroz, James; et. al. *Catálogo de clones de cacao recomendados por INIAP.* Yaguachi - Ecuador: Instituto Nacional de Investigaciones Agropecuarias - Estación Experimental Litoral Sur "Dr. Enrique Ampuero", 2014.

Ruiz Erazo, Ximena Andrea. *Diversidad genética de cacao Theobroma cacao L., con marcadores moleculares Microsatélites.* (tesis) (Maestría) Universidad Nacional de Colombia - Facultad de Ciencias Agrarias - Escuela de Posgrados, Palmira, Colombia. 2014.

Torcida, Sebastián, PEREZ, S. Iván. «Análisis de Procrustes y el estudio de la variación morfológica.» *Revista Argentina de Antropología Biológica* 14, nº 1 (2012): 131-141.

Vazquez Ovando, Alfredo; et al. «Potencial de los marcadores moleculares para el rescate de individuos de Theobroma cacao L. de alta calidad.» *BioTecnología* 16, nº 1 (2012): 36-56.

Villa Moyota, María Angélica. *Análisis Estadístico Multivariante de los Principales Componentes Químicos sobre Hachas Moneda de Cobre para determinar diferentes grupos de acuerdo a su composición.* (tesis) (Pre-grado) ESPOCH, Riobamba. 2012.

Xiong, R.; et al. «Permutation tests for Generalized Procrustes Analysis.» *Food Quality and Preference*, nº 19 (2008): 146-155.

Zuliani, Roxanna Paola. *Evaluación comparativa de las Técnicas Multivariadas Análisis Factorial Múltiple y Análisis de Procrustes Generalizado para el tratamiento de Datos de Tres Modos.* (tesis) (Maestría) Universidad Nacional de Córdoba, Argentina. 2012.